ignou
THE PEOPLE'S
UNIVERSITY

**School of Management Studies**

# MMPC-005
# QUANTITATIVE ANALYSIS FOR MANAGERIAL APPLICATIONS

**BLOCK 1 DATA COLLECTION AND ANALYSIS 5**
Unit 1 Collection of Data 7
Unit 2 Presentation of Data 18
Unit 3 Measures of Central Tendency 36
Unit 4 Measures of Variation and Skewness 57
**BLOCK 2 PROBABILITY AND PROBABILITY DISTRIBUTORS 75**
Unit 5 Basic Concepts of Probability 77
Unit 6 Discrete Probability Distributions 95
Unit 7 Continuous Probability Distributions 113
Unit 8 Decision Theory 131
**BLOCK 3 SAMPLING AND SAMPLING DISTRIBUTIONS 147**
Unit 9 Sampling Methods 149
Unit 10 Sampling Distributions 170
Unit 11 Testing of Hypotheses 192
Unit 12 Chi-Square Tests 224
**BLOCK 4 FORECASTING METHODS 249**
Unit13 Business Forecasting 251
Unit14 Correlation 268
Unit15 Regression 283
Unit 16 Time Series Analysis 308

ignou THE PEOPLE'S UNIVERSITY

# COURSE DESIGN AND PREPARATION TEAM

| | |
|---|---|
| Prof. K. Ravi Sankar,<br>Director, SOMS,<br>IGNOU, New Delhi | Prof. M. P. Gupta*<br>Faculty of Management Studies<br>University of Delhi |
| Dr. AshishChatterjee*<br>IIM, Calcutta | Dr. J. K. Sharma*<br>Faculty of Management Studies<br>University of Delhi |
| Prof. AbidHaleem<br>Faculty of Engineering and Technology,<br>JamiaMilliaIslamia,<br>New Delhi | Prof. P. K. Bhowmik*<br>International Management Institute<br>New Delhi |
| Prof. Kuldip Singh Sangwan<br>Mechanical Engineering Department,<br>Birla Institute of Technology and Science,<br>Pilani | Prof. H D Sharma<br>Former Prof & Head,<br>Pant Nagar Engineering College,<br>Pant Nagar |
| Prof. A. P. Verma<br>National Institute of Technology<br>Patna | Professor Ajay<br>Department of Industrial & Production Engineering,<br>G. B. Pant University of Agriculture & Technology, Pantnagar |
| Prof. Gokulananda Patel<br>Birla Institute of Management Technology<br>Greater Noida | Prof. Raj K Jain<br>Professor (Retd),Vikram University,<br>Ujjain |
| Prof. B.Sudheer<br>Dept of Management studies,<br>Sri Venkateswara University, Tirupati | Dr VSP Srivastav<br>Head (Retd), Computer Division,<br>IGNOU, New Delhi |
| ***Course Coordinator and Editor***<br>Prof. AnuragSaxena,<br>SOMS, IGNOU,<br>New Delhi | |

Note: A large portion of this course is adapted from the earlier MS-08 course and the persons marked with (*) are the original contributors of MS-8 Study Material. The profile of the expert given is as it was on the date of initial version.

# PRINT PRODUCTION

| | |
|---|---|
| Mr. Y.N. Sharma<br>Assistant Registrar<br>MPDD, IGNOU, New Delhi | Mr.Tilak Raj<br>Assistant Registrar<br>MPDD, IGNOU, New Delhi |

September, 2021

*© Indira Gandhi National Open University, 2021*

ISBN:

*All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University. Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at MaidanGarhi, New Delhi-110 068.*

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by the Registrar, MPDD, IGNOU.

Laser typeset by Tessa Media & Computers, C-206, A.F.E-II, Jamia Nagar, New Delhi-110025

ignou
THE PEOPLE'S
UNIVERSITY

# COURSE INTRODUCTION

This is a course which will introduce you to the basic concepts in quantitative techniques for managerial applications.

The **first unit** deals with sources, types, need and significance of data and data collection. The **second unit** systematically describes the classification and presentation of collected data.

The **third unit** gives an insight into treatment of data through central tendency measurement.

The **fourth unit** thoroughly discusses the deviations and different measures of variation.

The **fifth unit** gives you an insight into the concepts as such, different approaches, applications in different situations and their relevance in decision–making.

The **sixth and seventh units** deal with various application aspects of discrete and continuous probability distributions respectively in different situations.

The **eighth unit** systematically describes various approaches and analysis in decision theory enabling you to solve different decision problems.

The **ninth unit** deals with various aspects like rationale and types of sampling.

The **tenth unit** gives an insight into the concept of distribution and discusses the sampling distribution of some commonly used statistics.

The **eleventh unit** systematically describes the basic concepts of hypotheses, design, and use of tests concerning statistical hypotheses.

The **twelfth unit** gives you a clear understanding of the Chi-Square distribution and its role and significance in testing of hypotheses and decision making.

The **thirteenth unit** presents an overview of methods of business forecasting. Various methods suitable for long, medium and short term decisions are reviewed.

The **fourteenth unit** discusses the concept of correlation which is central in model development for forecasting. Various measures of the association between variables are described.

The **fifteenth unit** deals with a very important technique for establishing relationships between variables, namely regression. Fundamentals of linear regression are presented.

The **sixteenth unit** explains the basic concepts of time-series analysis. Here the objective is to forecast the future from the past by identifying the components like trend, seasonality, cyclic variations and randomness that may be present in historical data. An exposure to stochastic models is also given.

ignou THE PEOPLE'S UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# BLOCK 1

# DATA COLLECTION AND ANALYSIS

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 1 COLLECTION OF DATA

**Objectives**

- After studying this unit, you should be able to :
- Appreciate the need and significance of data collection
- Distinguish between primary and secondary data
- Know different methods of collecting primary data
- Design a suitable questionnaire
- Edit the primary data and know the sources of secondary data and its use
- Understand the concept of census vs. sample

**Structure**

1.1 Introduction
1.2 Primary and Secondary Data
1.3 Methods of Collecting Primary Data
1.4 Designing a Questionnaire
1.5 Pre-testing the Questionnaire
1.6 Editing Primary Data
1.7 Sources of Secondary Data
1.8 Precautions in the Use of Secondary Data
1.9 Census and Sample
1.10 Summary
1.11 Key Words
1.12 Self-assessment Exercises
1.13 Further Readings

ignou
THE PEOPLE'S
UNIVERSITY

## 1.1 INTRODUCTION

To make a decision in any business situation you need data. Facts expressed in quantitative form can be termed as data. Success of any statistical investigation depends on the availability of accurate and reliable data. These depend on the appropriateness of the method chosen for data collection. Therefore, data collection is a very basic activity in decision-making. In this unit, we shall be studying the different methods that are used for collecting data. Data may be classified either as primary or secondary.

## 1.2 PRIMARY AND SECONDARY DATA

Data used in statistical study is termed either "Primary" or "secondary" depending upon whether it was collected specifically for the study in question or for some other purpose. When the data used in a statistical study was collected under the control and supervision of the investigation, such type of data is referred to as “Primary data”. When the data was not collected by the investigator, but is derived from other sources then such data is referred to as “secondary data”.

The difference between primary and secondary data is only in terms of degree. For example, data which is primary in the hands of one become secondary in the hands of another. Suppose in investigator wants to study the working conditions of labour in a big industrial concerned. If he collects the data himself or through his agent, then this data is referred to as primary data. But if this data is used by someone else, then this data becomes secondary data.

## 1.3 METHOD OF COLLECTING PRIMARY DATA

Primary data may either be collected through the **observation method** or through the **questionnaire method.**

In the observation, the investigator asks no questions, but he simply observes the phenomenon under consideration, and records the necessary data. Sometimes individuals make the observation; on other occasion, mechanical and electronic devices do the job.

In the observation method, it may be difficult to produce accurate data. Physical difficulties on the part of the observer may result in errors. Because of these limitations in the observation method, the questionnaire method is most widely used for collecting data. In the questionnaire method, the investigator draws up a questionnaire containing all the relevant questions which he wants to ask from his respondents, and accordingly records the responses. Questionnaire method may be conducted through personal **interview**, or **by mail** or **telephone**.

**Personal Interviews** In this method the interviewer sits face-to-face with the respondent and records his responses. In this method, the information is likely to be more accurate and reliable because the interviewer can clear up doubts and cross-checks the respondents. This method is time-cons8uming and can be very costly if the number of respondents is large and widely distributed.

**Mail Questionnaire** In this method a list of questions (questionnaire) is prepare and mailed to the respondents. The respondents are expected to fill in the questionnaire and send it back to the investigator. Sometimes, mail questionnaire are placed in respondents’ hands through other means such as attaching them to consumers’ products or putting them in newspapers or magazines. This method can be easily adopted where the field of investigation is very vast and the respondents are spread over a wise geographical area. But this method can be adopted only where the respondents are literate and can understand written question and answer them.

**Telephone** In this method the investigator asks the relevant questions from the respondents over the telephone. This method is less expensive but it has limited application since only those respondents can be interviewed who have telephones; moreover, very few questions can be asked on telephone.

The questionnaire method is a very efficient and fast method of collecting data. But it has a very serious limitation as it may be extremely difficult to collect data on certain sensitive aspects such as income, age or personal life details, which the respondent may not be willing to share with the investigator. This is so with other methods also different people may interpret the questions differently and consequently there may be errors and inaccuracies in data collection.

**Activity A**

Explain clearly the observation and questionnaire methods of collecting primary data. Highlight their merits and limitation.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity B**

Describe the personal interviews and mail questionnaire method of data collection.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity C**

Point out the advantage of telephonic method of data collection. Does it have any limitations?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

Once the investigator has decided to use the questionnaire method the next step is to draw up a design of the survey.

ignou THE PEOPLE'S UNIVERSITY

A survey design involves the following steps :

a) Designing a questionnaire

b) Pre-testing a questionnaire

c) Editing the primary data.

## 1.4 DESIGNING OF QUESTIONNAIRE

The success of collecting data through a questionnaire depends mainly on how skillfully and imaginatively the questionnaire has been designed. A badly designed questionnaire will never be able to gather the relevant data. In designing the questionnaire, some of the important points to be kept in mind are:

**Covering letter** : Every questionnaire should be contain a covering letter. The covering letter should highlight the purpose of study and assure the respondent the all responses will be kept confidential. It is desirable that some inducement or motivation is provided to the respondent for better response. The objectives of the study and questionnaire design should be such that the respondent derives a sense of satisfaction through his involvement.

**Number of questions should be kept to the minimum:** The fewer the question, the greater the chances of getting a better responses and having all the questions answered. Otherwise the respondent may feel disinterested and provide inaccurate answers particularly towards the end of the questionnaire. Informing the question, the investigator has to take into consideration several factors such as the purpose of study, the time and resources available. As a rough indication, the number of questions should be between 15 to 40. In case the number of questions is more than 25, it is desirable that the questionnaire be divided into various part to ensure clarity.

**Questions should be simple, short and unambiguous:** The questions should be simple, short, easy to understand and such that their answers are unambiguous. For example, if the question is 'Are you literate? The respondent may have doubts about the meaning of literacy. To some literacy may mean a university degree whereas to others even the capacity to read and write may mean literacy. Hence it is desirable to specify whether you have passed (a) high school (b) graduation (c) post graduation etc. Questions can be of Yes/No type, or of multiple choice depending on the requirement o the investigator. Open-ended questions should generally be avoided.

**Questions of sensitive or personal nature should be avoided:** The questions should not be such as would require the respondent to disclose any private, personal or confidential information. For example, questions relating to sales, profits, material happiness etc. should be avoided as far as possible. If such questions are necessary in the survey, an assurance should be given to the respondent that the information provided shall be kept strictly confidential and shall not be used at any cost to their disadvantage.

**Answers to questions should not require calculations:** The questions should be framed in such a way that their answers do not require any calculations.

ignou THE PEOPLE'S UNIVERSITY

**Logical arrangement** The questions should be logically arranged so that there is a continuity of responses and the respondent does not feel the need to refer back to the previous questions. It is desirable that the questionnaire should begin with some introductory questions followed by vital questions crucial to he survey and ending with some light questions so that the overall impression of the respondent is a happy one.

**Cross-check and Footnotes:** The questionnaire should contain some such questions which act as a cross-check to the reliability of the information provided. For example, when a question relating to income is asked, it is desirable to include a question : "are you an income tax assessee?"

For the purpose of clarity, certain questions which might create a doubt in the mind of respondents, it is desirable to give footnotes. The purpose of footnotes is to clarify all possible doubts which may emerge from the questions and cannot be removed while answer them. For example, if a question relates to income limit like 1000-2000, 2000—3000; etc., a person getting exactly Rs. 2,000 should know in which income class he has to place himself.

One specimen format for a questionnaire used by IGNOU to elicit background of the participants and their expectations from the Diploma in Management course is shown below :

### INDIRA GANDHI NATIONAL OPEN UNIVERSITY
### SCHOOL OF MANAGEMENT STUDIES
### DIPLOMA IN MANAGEMENT
### OBJECTIVE – EXPECTATION ASSESSMENT FORMAT

A Name: …………………………………………………………………….

B Roll Number: …………………………………………………………..

C Name of your organization: ……………………………………………

D Nature of ownership of your Organisation (tick one)

[ ] Partnership
[ ] Public Ltd. Co.
[ ] Central Government
[ ] Autonomous
[ ] Private Limited Co.
[ ] Public Sector (Central/State)
[ ] Cooperative
[ ] Any other, Specify ……………….

E Designation ……………………………………………………………….

F What is your job level in the organizational hierarchy of your company?

Tick the appropriate box taking to Level as reference.

TOP LEVEL

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| [ ] | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] | [ ] |

If none specify _______________________________________________

G What is the nature of activities of your organization?

ignou THE PEOPLE'S UNIVERSITY

[ ] Manufacturing

[ ] Trading

[ ] Banking & Other Financial Services

[ ] Manufacturing

[ ] Professional & other services e.g. (hospital, P&T, Education etc.)

[ ] Civil Administration

[ ] Defence Services

[ ] Any other, specify

H What is the scale of operation of your organization?

I. *Value Turnover (in Rupees)*

[ ] Less than 50

[ ] More than 50 lakhs upto 1 crore

[ ] More than 1 crore upto 3 crore

[ ] More than 3 crore upto 7.5 crore

[ ] More than 12.5 crore upto 30 crore

[ ] More 30 crore (upto 50 crore)

[ ] More than 50 crore

*II. Number of Employees*

[ ] Less than 10

[ ] More than 10 upto 25

[ ] More than 25 upto 100

[ ] More than 100 upto 500

[ ] More than 500 upto 2000

[ ] More than 2000 upto 10000

[ ] More than 10000

I With what objectives have you joined this course?

State them in order of importance

………………………………………………………………………….

Most Important 1 ……………………………………………………..

2 ………………………………………………………

3 ………………………………………………………

4 ……………………………………………………..

5 ………………………………………………………

2 Would you employer appreciate and recognize your efforts in Doing this course? ( ) Yes ( ) No

If Yes, is he likely to reward you? ( ) Yes ( ) No

If Yes, state how ……………………………………………………….

3 How much time over and above the contact sessions, would you devote to studies for this programme *every week*?

[ ] Less than 2 hours

[ ] More than 2 hours upto 5 hours

[ ] More than 5 hours upto 10 hours

[ ] More than 10 hours

4 Have you had a chance to read the 3 blocks of print material sent to you this month?

Yes/No

If Yes, what do you like about them?

1) ……………………………..............................................................

ignou THE PEOPLE'S UNIVERSITY

2) ……………………………………………………………………………………

3) ……………………………………………………………………………………

And, what do you dislike about them?

1) ……………………………………………………………………………………

2) ……………………………………………………………………………………

3) ……………………………………………………………………………………

5 Which day(s) of the week is your office closed for weekly holiday(s) ………………………………………………………..

6 Give three preferences out of the following day and time slots for attending contact sessions. (1 = most preferred)

[ ] Monday 6.30 p.m. – 9.30 p.m.
[ ] Tuesday 6.30 p.m. – 9.30 p.m.
[ ] Wednesday 6.30 p.m. – 9.30 p.m.
[ ] Thursday 6.30 p.m. – 9.30 p.m.
[ ] Friday 6.30 p.m. – 9.30 p.m.
[ ] Saturday 10 a.m. – 1 p.m.
[ ] Saturday 6.30 p.m. 9.30 p.m.
[ ] Sunday 10 a.m. – 1. p.m
[ ] Sunday 6.30 p.m. -9.30 p.m.

**Activity D**

You have been directed by your employer to carry out a market survey to ascertain the probable demand for the new drug your company is going to introduce. Prepare a suitable questionnaire in this connection. State also the type of respondents you expect to cover.

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

## 1.5 PRE-TESTING THE QUESTIONNAIRE

Once the questionnaire has been designed, it is important to pre-test it. The pre-testing of a questionnaire is also known as pilot survey because it precedes the main survey work. Pre-testing allows rectification of problems, inconsistencies, repetitions etc. If changes are required, the necessary modifications can be made before administering the questionnaire, some questions are found irrelevant, they can be deleted and if some questions have to be included, the same can be done. Pre-testing must be done with utmost care, otherwise unnecessary and unwanted changes may be introduced. If time and resources permit, a second pre-testing can be also be done to ensure greater reliability of results. Proper testing, revising and re-testing would yield high dividends.

ignou THE PEOPLE'S UNIVERSITY

## 1.6 EDITING PRIMARY DATA

Once the questionnaires have been filled and the date collected, it is necessary to edit this data. Editing of data should be done to ensure completeness, consistency, accuracy and homogeneity.

**Completeness.** Each questionnaire should be complete in all respects, i.e. the respondent should have answered each and every question. If some important questions have been left unanswered, attempts should be made to contact the respondent and get the response. If despite al efforts, answered to vital questions are not given, such questionnaires should be dropped from final analysis.

**Consistency**. Questionnaire should also be checked to see that there are no contradictory answers. Contradictory responses my arise due to wrong answers filled up by te respondents or because of carelessness on the part of the investigator in recording the data. For example, the answers in a questionnaire to two successive question “Are you married?” and “Number of children you have?” may be given by a respondent as ‘No’ and ‘Two’ respectively. Obviously, there is some inconsistency in the answers to these two questions which should be sorted out with the respondent.

**Accuracy**. The questionnaire should also be checked for the accuracy of information provided by the respondent. It may be pointed out that this is the most difficult job of the investigator and at the same time the most important one. If inaccuracies are permitted, this would lead to misleading results. Inaccuracies may be checked by random cross-chekcing.

**Homogeneity.** It is equally important to check whether the questions have been understood in the same sense by all the respondents. For instance, if there is a question on income, it should be very clearly stated whether it refers to weekly, monthly, or yearly income. If it is left ambiguous then respondents may give different responses and there will be no basis for comparison because we may take some figures which are valid for monthly income and some for annual income.

## 1.7 SOURCES OF SECONDARY DATA

The sources of secondary data may be divided into two broad categories, published and unpublished.

**Published Sources.** There are a number of national and international organizations which collect statistical data and publish their findings in statistical reports periodically. Some of the national organizations which collect, compile and publish statistical data are : Central Statistical Organization (CSO); National Sample Survey Organization (NSSO); Office of the Registrar General and Census Commissioner of India; Labour Bureau; Federation of Indian Chambers of Commerce and Industry; Indian Council of Agricultural Research (ICAR); The Economic Times; The Financial Express etc. Some of the international agencies which provide valuable statistical data on a variety of social-economic and political events are : United Nations

Organization (UNO); World Health Organization (WHO); International Labour Organization (ILO); International Monetary Fund (IMF); World Bank etc.

**Unpublished Sources.** All statistical data need no the published. A major sources of statistical data produced by government, semi-government , private and public organizations is based on the date drawn from internal records. This data based on internal records provides authentic statistical data and is much cheaper as compared to primary data. Some example of the internal records include employees' payroll, the amount of raw materials, cash receipts and cash book etc. It may be pointed out that it is very difficult to have access to unpublished information.

## 1.8 PRECAUTIONS IN THE USE OF SECONDARY DATA

A careful scrutiny must be made before using published data. The user should be extra cautious in using secondary data and he should not accept it at its face value. The reason may be that such data is full of errors because of bias, inadequate sample size, errors of definitions and computational errors etc. Therefore, before using such data, the following aspects should be considered.

**Suitability.** The investigator mush ensure that the data available is suitable for the purpose of the inquiry on hand. The suitability of data may be judged by comparing the nature and scope of investigation.

**Reliability.** It is of utmost importance to determine how reliable is the data from secondary source and how confidently we can use it. In assessing the reliability, it is important to know whether the collecting agency is unbiased, whether it has a representative sample the data whether has been properly analyzed, as so on.

**Adequacy**. Data from secondary sources may be available but its scope may be limited and therefore this may not serve the purpose of investigation. The data may cover only a part of the requirement of the investigator or may pertain to a different time period.

Only if the investigator is fully satisfied on all the above mentioned points, he should proceed with this data a the starting point for further analysis.

## 1.9 CENSUS AND SAMPLE

When secondary data is not available for the problem under study, a decision may be taken to collect primary data through original investigation. This original investigation may be obtained either by census (or complete enumeration) method or sampling method. When the investigator collects data about each and every item in the population, it is known as the census method or complete enumeration survey. But when the investigator studies only a representative part of the total population and makes inferences about the population on the basis of that study, it is known as the sampling method.

In both the situations, the investigator is interested in studying some characteristics of the population.

The advantage of the census method is that information about every item in the population can be obtained. Also the information collected is more accurate. The main limitations of the census method are that it requires a great deal of money and time. Moreover in certain practical situations of quality control, such as finding the tensile strength of a steel specimen by stretching it till it breaks is not even physically possible to check each and every item because quality testing result in the destruction of the item itself. In most cases, it is not necessary to study every unit of the population to draw some inference about. If a sample is representative of the population then our study of the sample will yield correct inference about the total population.

It should be noted that out of the census and sampling methods, the sampling method is much more widely used in practice. There are several methods of sampling which would be discussed in detail in nit 13 on ‘sampling methods’.

## 1.10 SUMMARY

Statistical data is a set of facts expressed in quantitative form. The use of facts expressed as measurable quantities can help a decision maker to arrive at better decisions. Data can be obtained through primary sources or secondary source. When the data is collected by the investigator himself, it is called primary data. When the data has been collected by others it is known as secondary data. The most important method for primary data collection is through questionnaire. A questionnaire refers to a device used to secure answers to questions from the respondents. Another important distinction in considering data is whether the values represent the complete enumeration of some whole, known as population or universe, or only a part of the population, which is called a sample.

## 1.11 KEY WORDS

**Census** is the collection of each and every item in the given population or universe.

**Population** is the collection of items on which information is required

**Primary Data** is the collection of data by the investigator himself.

**Questionnaire** is a device for getting answers to questions by using a form to which the respondent responds.

**Sample** is any group of measurements selected form a population.

**Secondary Data** is the collection of data compiled by someone other than the user.

## 1.12 SELF-ASSESSMENT EXERCISES

1. Distinguish between primary and secondary data. Discuss the various methods of collecting primary data. Indicate the situation in which each of these methods should be used.
2. Discuss the validity of the statement : “A secondary source is not as reliable as a primary source”.
3. Discuss the various sources of secondary data. Point out the precautions to be taken while using such data.
4. Describe briefly the questionnaire method of collecting primary data. State the essentials of a good questionnaire.
5. Explain what precautions must be taken while drafting a useful questionnaire.
6. As the personnel manager in a particular industry, you are asked to deter4mine the effect of increased wages on output. Draft a suitable questionnaire for this purpose.
7. If you were to conduct a survey regarding smoking habits among students of IGNOU, what method of data collection would you adopt? Give reasons for your choice.
8. Distinguish between the census and sampling methods of data collections and compare their merits and demerits. Why is the sampling method unavoidable in certain situation?
9. Explain the terms ‘Population’ and ‘sample’. Explain why it is sometimes necessary and often desirable to collect information about the population by conducting a sample survey instead of complete enumeration.

## 1.13 FURTHER READINGS

Clark, T.C. and E.W. Jordan. *Introduction to Business and Economic Statistics,* South-Western Publising Co.: Ohio.

Enns, P.G. Business Statistics, Richard D. Irwin Inc.: Homewood.

Gupta, S.P. and M.P. Gupta. Business Statistics, Sultan Chand & Sons: New Delhi

Levin, R.I. Statistics for Management, Prentice Hall of India: New Delhi.

Moskowitz, H. and G.P. Wright. Statistics for Management and Economics, Charles E. Merill Publishing Company : Ohio

# UNIT 2 PRESENTATION OF DATA

**Objectives**

After studying this unit, you should be able to :

- understand the need and significance of presentation of data
- know the necessity of classifying data and various types of classification
- construct a frequency distribution of discrete and continuous data
- present a frequency distribution in the form of bar diagram, histogram, frequency polygon, and ogives.

**Structure**

2.1 Introduction
2.2 Classification of Data
2.3 Objectives of Classification
2.4 Types of Classification
2.5 Construction of a Discrete Frequency Distribution
2.6 Construction of a Continuous Frequency Distribution
2.7 Guidelines for Choosing the Classes
2.8 Cumulative and Relative Frequencies
2.9 Charting of Data
2.10 Summary
2.11 Key Words
2.12 Self-assessment Exercises
2.13 Further Readings

## 2.1 INTRODUCTION

In the previous unit, we discussed the various ways of collecting data. The successful use of the data collected depends to a great extent upon the manner in which it is arranged, displayed and summarised. In this unit, we shall be mainly interested in the presentation of data. Presentation of data can be displayed either in tabular form or through charts. In the tabular form, it is necessary to classify the data before the data is tabulated. Therefore, this unit is divided into two section, viz., (a) classification of data and (b) charting of data.

## 2.2 CLASSIFICATION OF DATA

After the data has been systematically collected and edited, the first step in presentation of data is classification. Classification is the process of arranging the data according to the points of similarities and dissimilarities. It is like the process of sorting the mail in a post office where the mail for different destinations is placed in different compartments after it has been carefully sorted cut from the huge heap.

## 2.3 OBJECTIVES OF CLASSIFICATION

The principal objectives of classifying data are:

- to condense the mass of data in such a way that salient features can be readily noticed
- to facilitate comparisons between attributes of variables
- to prepare data which can be presented in tabular form
- to highlight the significant features of the data at a glance

## 2.4 TYPES OF CLASSIFICATION

Some common types of classification are:

1) Geographical i.e., according to area or region.
2) Chronological, i.e., according to occurrence of an event with respect to time.
3) Qualitative, i.e., according to attributes.
4) Quantitative, i.e., according to magnitudes.

**Geographical Classification**. In this type of classification, data is classified according to area or region. For example, when we consider production of wheat statewise, this would be called geographical classification. The listing of individual entries are generally done in an alphabetical order or according to size to emphasise the importance of a particular area or region.

Chronological Classification. When the data is classified according to the time of its occurrence, it is known as chronological classification. For example, sales figure of a company for last six years are given below:

| Year | Sales (Rs. lakhs) | Year | Sales (Rs. Iakhs) |
|---|---|---|---|
| 1982-83 | 175 | 1985-86 | 485 |
| 1983-84 | 220 | 1986-87 | 565 |
| 1984-85 | 350 | 1987-88 | 620 |

**Qualitative Classification.** When the data is classified according to some attributes (distinct categories) which are not capable of measurement is known as qualitative classification. In a simple (or dichotomous) classification, an attribute is divided into two classes, one possessing the attribute and the other not possessing it. For example, we may classify population on the basis of employment, i.e., the employed and the unemployed. Similarly, we can have manifold classification when an attribute is divided so as to form several classes. For example, the attribute education can have different classes such as primary, middle, higher secondary, university, etc.

**Quantitative Classification.** When the data is classified according to some characteristics that can be measured, it is called quantitative classification. For example, the employees of a company may be classified according to their monthly salaries. Since quantitative data is characterised by different numerical values, the data represents the values of a variable. Quantitative data may be further classified into one or two types: discrete or continuous.

The term discrete data refers to quantitative data that is limited to integer numerical values of a variable. For example, the number of employees in an organisation or the number of machines in a factory are examples of discrete data.

Continuous data can take integer as well as fraction values of the variable. For example, the data relating to weight, distance, and volume are examples of continuous data. The quantitative classification becomes the basis for frequency distribution.

When the data is arranged into groups or categories according to conveniently established divisions of the range of the observations, such an arrangement in tabular form is called a frequency distribution. In a frequency distribution, raw data is represented by distinct groups which are known as **classes**. The number of observations that fall into each of the classes is known as **frequency**. Thus, a frequency distribution has two parts, on its left there are classes and on its right there are frequencies.

When data is described by a continuous variable it is called continuous data and when it is described by a discrete variable, it is called **discrete data**. The following are the two examples of discrete and continuous frequency distributions.

| **Discrete frequency distribution** | | **Continuous frequency distribution** | |
|---|---|---|---|
| No. of employees | No. of companies | Age (Years) | No. of workers |
| 110 | 25 | 20-25 | 15 |
| 120 | 35 | 25-30 | 22 |
| 130 | 70 | 30-35 | 38 |
| 140 | 100 | 35-40 | 47 |
| 150 | 18 | 40-45 | 18 |
| 160 | 12 | 45-50 | 10 |

**Activity A**

What do you understand by classification of data?

Why classification is necessary?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity B**

With the help of a suitable example, illustrate the difference between qualitative and quantitative data.

…………………………………………………………………………………….

…………………………………………………………………………………….

…………………………………………………………………………………….

…………………………………………………………………………………….

## 2.5 CONSTRUCTION OF A DISCRETE FREQUENCY DISTRIBUTION

The process of preparing a frequency distribution is very simple. In the case discrete data, place all possible values of the variable in ascending order in column, and then prepare another column of 'Tally' marks to count the number of times a particular value of the variable is repeated. To facilitate counting, block of five 'Tally' marks are prepared and some space is left in between the blocks. The frequency column refers to the number of 'Tally' marks, a particular class will contain. To illustrate the construction of a discrete frequency distribution, consider a sample study in which 50 families were surveyed to find the number of children per family. The data obtained are:

3 2 2 1 3 4 2 1 3 4 5 0 2
1 2 3 3 2 1 1 2 3 0 3 2 1
4 3 5 5 4 3 6 5 4 3 1 0 6
5 4 3 1 2 0 1 2 3 4 5

To condense this data into a discrete frequency distribution, we shall take the help of 'Tally' marks as shown below:

| No. of Children | No. of families | Frequency |
|---|---|---|
| 0 | IIII | 4 |
| 1 | 𝍸 IIII | 9 |
| 2 | 𝍸 𝍸 | 10 |
| 3 | 𝍸 𝍸 II | 12 |
| 4 | 𝍸 II | 7 |
| 5 | 𝍸 I | 6 |
| 6 | II | 2 |
| | | Total 50 |

## 2.6 CONSTRUCTION OF A CONTINUOUS FREQUENCY DISTRIBUTION

In constructing the frequency distribution for continuous data, it is necessary to clarify some of the important terms that are frequently used.

ignou THE PEOPLE'S UNIVERSITY

**Class Limits.** Class limits denote the lowest and highest value that can be included in the class. The two boundaries (i.e., lowest and highest) of a class are known as the lower limit and the upper limit of the class. For example, in the class 60-69, 60 is the lower limit and 69 is the upper limit or we can say that there can be no value if that class which is less than 60 and more than 69.

**Class Intervals.** The class interval represents the width (span or size) of a class. The width may be determined by subtracting the lower limit of one class from the lower limit of the following class (alternatively successive upper limits may be used). For example, if the two classes are 10-20 and 20-30, the width of the class interval would be the difference between the two successive lower limits, i.e., 20-10 = 10 or the difference between the upper limit and lower limit of the same class, i.e., 20-10 = 10.

**Class Frequency.** The number of observations falling within a particular class is called its class frequency or simply frequency. Total frequency (sum of all the frequencies) indicate the total number of observations considered in a given frequency distribution.

**Class Mid-point.** Mid-point of a class is defined as the sum of class limits divided by 2. Therefore, it is the value lying halfway between the lower and upper class limits. In the example taken above the mid-point would be (10+20)/2 = 15 corresponding to the class 10-20 and 25 corresponding to the class 20-30.

**Type of Class Interval.** There are different ways in which limits of class intervals can be shown such as:

i) Exclusive and Inclusive method, and

ii) Open-end

**Exclusive Method.** The class intervals are so arranged that the upper limit of one class is the lower limit of the next class. The following example illustrates this point.

| Sales (Rs. thousands) | No. of firms | Sales (Rs. thousands) | No. of firms |
|---|---|---|---|
| 20-25 | 20 | 35-40 | 27 |
| 25-30 | 28 | 40-45 | 12 |
| 30-35 | 35 | 45-50 | 8 |

In the above example there are 20 firms whose sales are between Rs.20,000 and Rs. 24,999. A firm with sales of exactly Rs. 25 thousand would be included in the next class viz. 25-30. Therefore in the exclusive method, it is always presumed that upper limit is excluded.

**Inclusive Method.** In this method, the upper limit of one class is included in that class itself. The following example illustrates this point.

| Sales (Rs. thousands) | No. of firms | Sales (Rs. thousands) | No. of firms |
|---|---|---|---|
| 20-24.999 | 20 | 35-39.999 | 27 |
| 25-29.999 | 28 | 40-44.999 | 12 |
| 30-34.999 | 35 | 45-49.999 | 8 |

In this example, there are 20 firms whose sales are between Rs. 20,000 and Rs. 24,999. A firm whose sales are exactly Rs. 25,000 would be included in the next class. Therefore in the inclusive method, it is presumed that upper limit is included.

It may be observed that both the methods give the same class frequencies, although the class intervals look different. Whenever inclusive method is used for equal class intervals, the width of class intervals can be obtained by taking the difference between the two lower limits (or upper limits).

**Open-End.** In an open-end distribution, the lower limit of the very first class and upper limit of the last class is not given. In distribution where there is a big gap between minimum and maximum values, the open-end distribution can be used such as in income distributions. The income disparities, of residents of a region may vary between Rs. 800 to Rs. 50,000 per month. In such a case, we can form classes like:

Less than Rs. 1,000
1,000-2,000
2,000-5,000
5,000-10,000
10,000-25,000
25,000 and above

**Remark.** To ensure continuity and to get correct class intervals, we shall adopt exclusive method. However, if inclusive method is suggested then it is necessary to make an adjustment to determine the class interval. This can be done by taking the average value of the difference between the lower limit of the succeeding class and the upper limit of the class. In terms of formula:

$$\text{Correction factor} = \frac{\text{Lower Limit of second class} - \text{Upper Limit of the first class}}{2}$$

This value so obtained is deducted from all lower limits and added to all upper limits. For instance, the example discussed for inclusive method can easily be converted into exclusive case. Take the difference between 25 and 24,999 and divide it by 2. Thus correction factor becomes (25-24,999)/2 = 0.0005. Deduct this value from lower limits and add it to upper limits. The new frequency distribution will take the following form:

Presentation of Data

| Sales (Rs thousand) | No. of firms | Sales (Rs thousand) | No. of firms |
|---|---|---|---|
| 19.9995-24.9995 | 20 | 34.9995-39.9995 | 27 |
| 24 9995-29 9995 | 28 | 39 9995-44 9995 | 12 |
| 29.9995-34.9995 | 35 | 44.9995-49.9995 | 8 |

ignou
THE PEOPLE'S
UNIVERSITY

## 2.7 GUIDELINES FOR CHOOSING THE CLASSES

The following guidelines are useful in choosing the class intervals.

1) The number of classes should not be too small or too large. Preferably, the number of classes should be between 5 and 15. However, there is no hard and fast rule about it. If the number of observations is smaller, the number of classes formed should be towards the lower side of this limit and when the number of observations increase, the number of classes formed should be towards the upper side of the limit.
2) If possible, the widths of the intervals should be numerically simple like 5, 10, 25 etc. Values like 3, 7, 19 etc. should be avoided.
3) It is desirable to have classes of equal width. However, in case of distributions having wide gap between the minimum and maximum values, classes with unequal class interval can be formed like income distribution.
4) The starting point of a class should begin with 0, 5, 10 or multiples thereof. For example, if the minimum value is 3 and we are taking a class interval of 10, the first class should be 0-10 and not 3-13.
5) The class interval should be determined after taking into consideration the minimum and maximum values and the number of classes to be formed. For example, if the income of 20 employees in a company varies between Rs. 1100 and Rs. 5900 and we want to form 5 classes, the class interval should be 1000

$$\frac{5900\ -\ 1100}{1000} = 4.8\ or\ 5$$

All the above points can be explained with the help of the following example wherein the ages of 50 employees are given:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 21 | 37 | 33 | 28 | 42 | 56 | 33 | 32 | 59 |
| 40 | 47 | 29 | 65 | 45 | 48 | 55 | 43 | 42 | 40 |
| 37 | 39 | 56 | 54 | 38 | 49 | 60 | 37 | 28 | 27 |
| 32 | 33 | 47 | 36 | 35 | 42 | 43 | 55 | 53 | 48 |
| 29 | 30 | 32 | 37 | 43 | 54 | 55 | 47 | 38 | 62 |

In order to form the frequency distribution of this data, we take the difference between 60 and 21 and divide it by 10 to form 5 classes as follows:

| Age (Years) | Tally Marks | Frequency |
|---|---|---|
| 20-30 | 卌 II | 7 |
| 30-40 | 卌 卌 卌 I | 16 |
| 40-50 | 卌 卌 卌 | 15 |
| 50-60 | 卌 IIII | 9 |
| 60-70 | IIII | 3 |
| | | Total 50 |

**Activity C**

Distinguish between the following:

i) Discrete and continuous frequency distributions.

ii) Class limits and class intervals.

iii) Inclusive and Exclusive method.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 2.8 CUMULATIVE AND RELATIVE FREQUENCIES

It is often useful to express class frequencies in different ways. Rather than listing the actual frequency opposite each class, it may be appropriate to list either cumulative, frequencies or relative frequencies or both.

**Cumulative Frequencies.** As its name indicates, it cumulates the, frequencies, starting at either the lowest or highest value. The cumulative frequency of a given class interval thus represents the total of all the previous class frequencies including the class against which it is written. To illustrate the concept of cumulative frequencies, consider the following example

| Monthly salary (Rs.) | No. of employees | Monthly salary (Rs.) | No. of employees |
|---|---|---|---|
| 1000-1200 | 5 | 2000-2200 | 25 |
| 1200-1400 | 14 | 2200-2400 | 22 |
| 1400-1600 | 23 | 2400-2600 | 7 |
| 1600-1800 | 50 | 2600-2800 | 2 |
| 1800-2000 | 52 | | |

If we keep on adding the successive frequency of each class starting from the frequency of the very first class, we shall get cumulative frequencies as shown below:

| Monthly salary (Rs.) | No. of employees | Cumulative |
|---|---|---|
| 1000-1200 | 5 | 5 |
| 1200-1400 | 14 | 19 |
| 1400-1600 | 23 | 42 |
| 1600-1800 | 50 | 92 |
| 1800-2000 | 52 | 144 |
| 2000-2200 | 25 | 169 |
| 2200-2400 | 22 | 191 |
| 2400-2600 | 7 | 198 |
| 2600-2800 | 2 | 200 |
|  | Total 200 |  |

**Relative Frequencies.** Very often, the frequencies in a frequency distribution are converted to relative frequencies to show the percentage for each class. If the frequency of each class is divided by the total number of observations (total frequency), then this proportion is referred to as relative frequency. To get the percentage for each class, multiply the relative frequency by 100; For the above example, the values computed for relative frequency and percentage are shown below:

| Monthly salary (Rs.) | No. of employees | Relative frequency | Percentage |
|---|---|---|---|
| 1000-1200 | 5 | 0.025 | 2.5 |
| 1200-1400 | 14 | 0.070 | 7.0 |
| 1400-1600 | 23 | 0.115 | 11.5 |
| 1600-1800 | 50 | 0.250 | 25.0 |
| 1800-2000 | 52 | 0.260 | 26.0 |
| 2000-2200 | 25 | 0.125 | 12.5 |
| 2200-2400 | 22 | 0.110 | 11.0 |
| 2400-2600 | 7 | 0.035 | 3.5 |
| 2600-2800 | 2 | 0.010 | 1.0 |
|  | 200 | 1.000 | 100% |

There are two important advantages in looking at relative frequencies (percentages) instead of absolute frequencies in a frequency distribution.

1) Relative frequencies facilitate the comparisons of two or more than two sets of data.

2) Relative frequencies constitute the basis of understanding the concept of probability.

**Activity D**

With the help of an example, explain the concept of relative frequency.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 2.9 CHARTING OF DATA

Charts of frequency distributions which cover both diagrams and graphs are useful because they enable a quick interpretation of the data. A frequency distribution can be presented by a variety of methods. In this section, the following four popular methods of charting frequency distribution are discussed in detail.

i) Bar Diagram

ii) Histogram

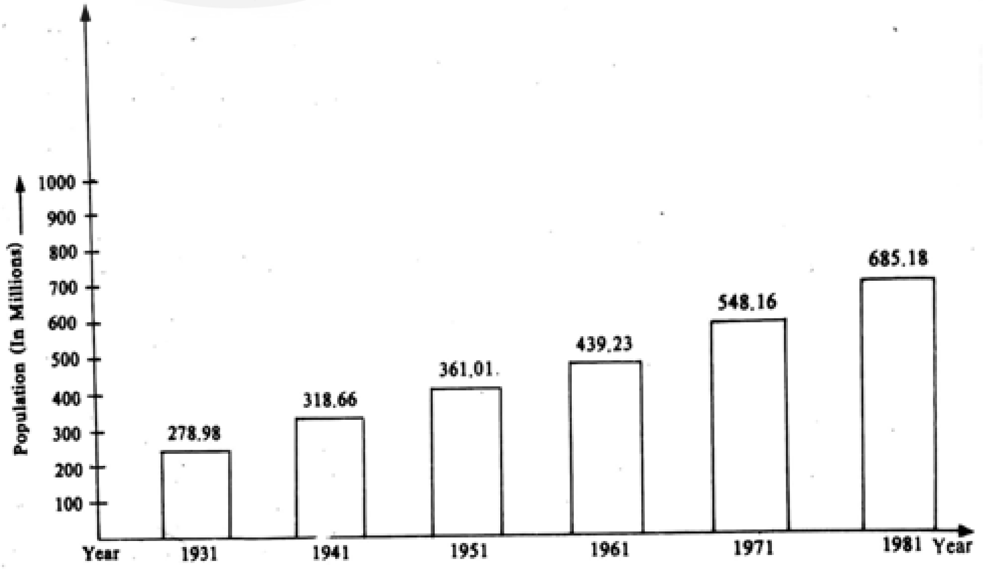
iii) Frequency Polygon

iv) Ogive or Cumulative Frequency Curve

**Bar Diagram.** Bar diagrams are most popular. One can see numerous such diagrams in newspapers, journals, exhibitions, and even on television to depict different characteristics of data. For example, population, per capita income, sales and profits of a company can be shown easily through bar diagrams. It may be noted that a bar is a thick line whose width is shown to attract the viewer. A bar diagram may be either vertical or horizontal.

In order to draw a bar diagram, we take the characteristic (or attribute) under consideration on the X-axis and the corresponding value on the Y-axis. It is desirable to mention the value depicted by the bar on the top of the bar.

To explain the procedure of drawing a bar diagram, we have taken the population figures (in millions) of India which are given below:



Bar Diagram

ignou THE PEOPLE'S UNIVERSITY

Take the years on the X-axis and the population figure on the Y-axis and draw a bar to show the population figure for the particular year. As can be seen from the diagram, the gap between one bar and the other bar is kept equal. Also the width of different bars is same. The only difference is in the length of the bars and that is why this type of diagram is also known as one dimensional.

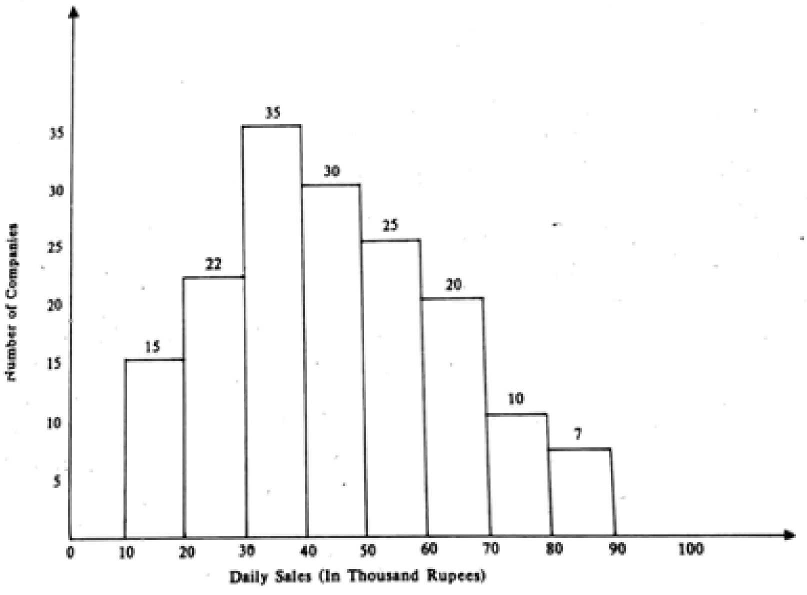
**Histogram.** One of the most commonly used and easily understood methods for graphic presentation of frequency distribution is histogram. A histogram is a series of rectangles having areas that are in the same proportion as the frequencies of a frequency distribution.

To construct a histogram, on the horizontal axis or X-axis, we take the class limits of the variable and on the vertical axis or Y-axis, we take the frequencies of the class intervals shown on the horizontal axis. If the class intervals are of equal width, then the vertical bars in the histogram are also of equal width. On the other hand, if the class intervals are unequal, then the frequencies have to be adjusted according to the width of the class interval. To illustrate a histogram when class intervals are equal, let us consider the following example.

| Daily sales (Rs. thousand) | No. of companies | Daily sales (Rs. thousand) | No. of companies |
|---|---|---|---|
| 10-20 | 15 | 50-60 | 25 |
| 20-30 | 22 | 60-70 | 20 |
| 30-40 | 35 | 70-80 | 16 |
| 40-50 | 30 | 80-90 | 7 |

In this example, we may observe that class intervals are of equal width. Let us take class intervals on the X-axis and their corresponding frequencies on the Y-axis. On each class interval (as base), erect a rectangle with height equal to the frequency of that class. In this manner we get a series of rectangles each having a class interval as its width and the frequency as its height as shown below:

**Histogram with Equal Class Intervals**



It should be noted that the area of the histogram represents the total frequency as distributed throughout the different classes.

When the width of the class intervals are not equal, then the frequencies must be adjusted before constructing the histogram.

The following example will illustrate the procedure:

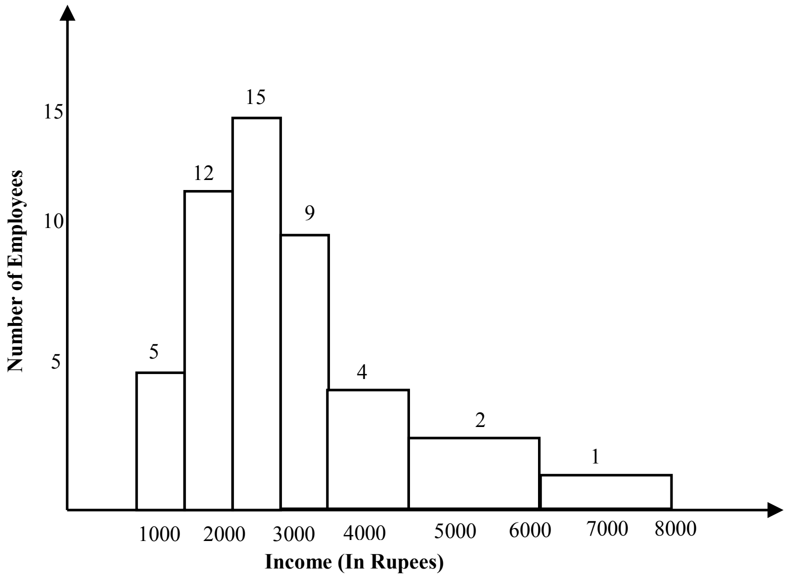
| Income (Rs.) | No. of employees | Income (Rs.) | No. of |
|---|---|---|---|
| 1000-1500 | 5 | 3500-5000 | 12 |
| 1500-2000 | 12 | 5000-7000 | 8 |
| 2000-2500 | 15 | 7000-8000 | 2 |
| 2500-3500 | 18 | | |

As can be seen, in the above example, the class intervals are of unequal width and hence we have to find out the adjusted frequency of each class by taking the class with the lowest class interval as the basis of adjustment. For example, in the class 2500-3500, the class interval is 1000 which is twice the size of the lowest class interval, i.e., 500 and therefore the frequency of this class would be divided by two, i.e., it would be 18/2 = 9. In a similar manner, the other frequencies would be obtained. The adjusted frequencies for various classes are given below:

| Income (Rs.) | No. of employees | Income (Rs.) . | No. of employees |
|---|---|---|---|
| 1000-1500 | 5 | 4000-4500 | 4 |
| 1500-2000 | 12 | 4500-5000 | 4 |
| 2000-2500 | 15 | 5000-5500 | 2 |
| 2500-3000 | 9 | 5500-6000 | 2 |
| 3000-3500 | 9 | 6000-6500 | 2 |
| 3500-4000 | 4 | 6500-7000 | 2 |
| | | 7000-7500 | 1 |
| | | 7500-8000 | 1 |

The histogram of the above distribution is shown below:

**Histogram with Unequal Class Intervals**



ignou
THE PEOPLE'S
UNIVERSITY

It may be noted that a histogram and a bar diagram look very much alike but have distinct features. For example, in a histogram, the rectangles are adjoining and can be of different width whereas in bar diagram it is not possible.

**Activity E**

Draw a sketch of a histogram and a bar diagram and explain the difference between the two.

………………………………………………………………………………………

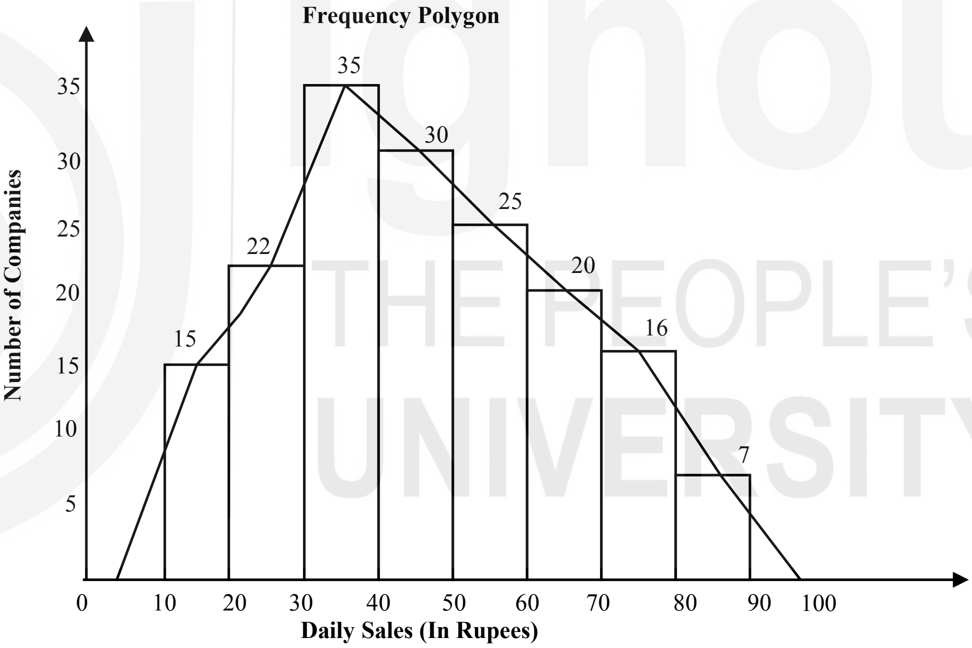
………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

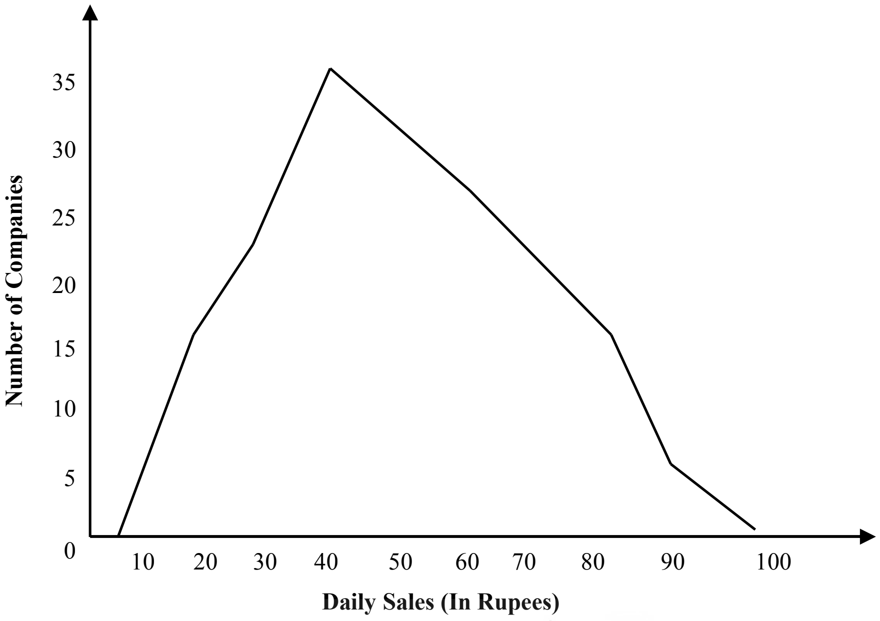
**Frequency Polygon.** The frequency polygon is a graphical presentation of frequency distribution. A polygon is a many sided closed figure.



A frequency polygon is constructed by taking the mid-points of the upper horizontal side of each rectangle on the histogram and connecting these mid-points by straight lines. In order to close the polygon, an additional class is assumed at each end, having a zero frequency.

If we draw a smooth curve over these points in such a way that the area included under the curve is approximately the same as that of the polygon, then such a curve is known as frequency curve. The following figure shows the same data smoothed out to form a frequency curve, which is another form of presenting the same data.

Frequency Curve



**Remark.** The histogram is usually associated with discrete data and a frequency polygon is appropriate for continuous data. But this distinction is not always followed in practice and many factors may influence the choice of graph.

The frequency polygon and frequency curve have a special advantage over the histogram particularly when we want to compare two or more frequency distributions.

**Activity F**

What is the procedure of making a frequency polygon?

Illustrate with the help of suitable data.

………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

**Ogives or Cumulative Frequency Curve.** An ogive is the graphical presentation of a cumulative frequency distribution and therefore when the graph of such a distribution is drawn, it is called cumulative frequency curve or ogive. There are two methods of constructing ogive; viz.,

i) Less than ogive

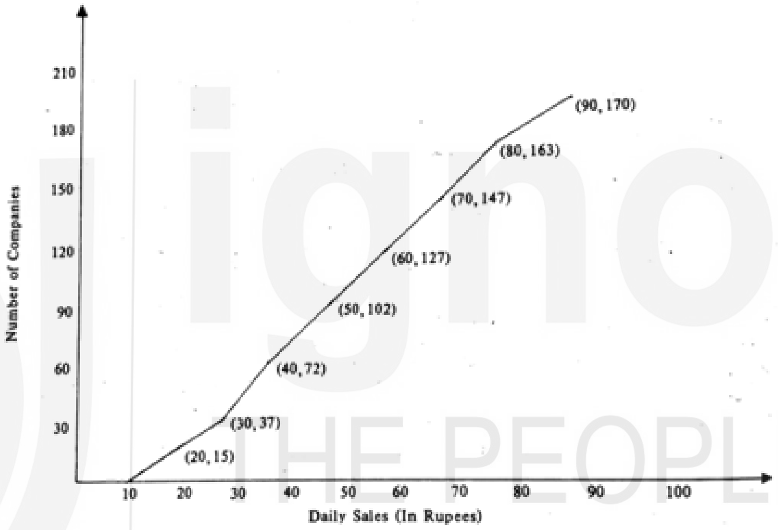
ii) More than ogive

**Less than Ogive.** In this method, the upper limit of the various classes are taken on the X-axis and the frequencies obtained by the process of cumulating the preceding frequencies on the Y-axis. By joining these points we get less than ogive. Consider the example relating to daily sales discussed earlier.

| Daily sales (Rs. thousand) | No. of companies | Daily sales (Rs. thousand) | No. of companies |
|---|---|---|---|
| 10-20 | 15 | Less than 20 | 15 |
| 20-30 | 22 | Less than 30 | 37 |
| 30-40 | 35 | Less than 40 | 72 |
| 40-50 | 30 | Less than 50 | 102 |
| 50-60 | 25 | Less than 60 | 127 |
| 60-70 | 20 | Less than 70 | 147 |
| 70-80 | 16 | Less than 80 | 163 |
| 80-90 | 7 | Less than 90 | 170 |

The less than Ogive Curve is shown below:

**Less than Ogive**



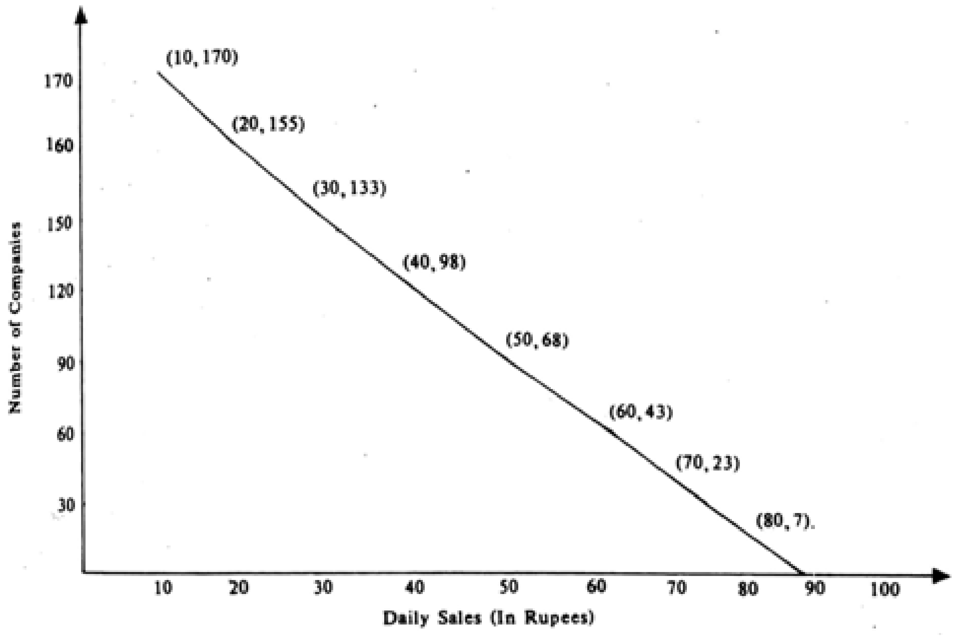
**More than Ogive.**Similarly more than ogive or cumulative frequency curve can be drawn by taking the lower limits on X-axis and cumulative frequencies on the Y-axis. By joining these points, we get more than ogive. The table and the curve for this case is shown below:

| Daily sales (Rs, thousand) | No. of companies | Daily sales (Rs. thousand) | Cumulative frequency |
|---|---|---|---|
| 10-20 | 15 | More than 10 | 170 |
| 20-30 | 22 | More than 20 | 155 |
| 30-40 | 35 | More than 30 | 133 |
| 40-50 | 30 | More than 40 | 98 |
| 50-60 | 25 | More than 50 | 68 |
| 60-70 | 20 | More than 60 | 43 |
| 70-80 | 16 | More than 70 | 23 |
| 80-90 | 7 | More than 80 | 7 |

The more than ogive curve is shown below:

**More than Ogive**

ignou
THE PEOPLE'S
UNIVERSITY



The shape of less than ogive curve would be a rising one whereas the shape of more than ogive curve should be falling one.

The concept of ogive is useful in answering questions such as: How many companies are having sales less than Rs. 52,000 per day or more than Rs. 24,000 per day or between Rs. 24,000 and Rs. 52,000?

**Activity G**

With the help of an example, explain the concept of less than ogive and more than ogive.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 2.10 SUMMARY

Presentation of data is provided through tables and charts. A frequency distribution is the principal tabular summary of either discrete or continuous data. The frequency distribution may show actual, relative or cumulative frequencies. Actual and relative frequencies may be charted as either histogram (a bar chart) or a frequency polygon. Two graphs of cumulative frequencies are: less than ogive or more than ogive.

## 2.11 KEY WORDS

**Bar Chart** is a thick line where the length of the bars should be proportional to the magnitude of the variable they present.

**Class Interval** represents the width of a class.

ignou
THE PEOPLE'S
UNIVERSITY

**Class Limits** denote the lowest and highest value that can be included in the class.

**Continuous Data** can take all values of the variable.

**Discrete Data** refers to quantitative data that are limited to non-negative integral numerical values of a variable.

**Frequency Distribution** is a tabular presentation where a number of observations with similar or closely related values are put in groups.

**Qualitative Data** is characterised by exhaustive and distinct categories that do not possess magnitude.

**Quantitative Data** possess the characteristic of numerical magnitude.

## 2.12 SELF-ASSESSMENT EXERCISES

1) Explain the purpose and methods of classification of data giving suitable examples.
2) What are the general guidelines of forming a frequency distribution with particular reference to the choice of class intervals and number of classes?
3) Explain the various diagrams and graphs that can be used for charting a frequency distribution.
4) What are ogives? Point out the role. Discuss the method of constructing ogives with the help of an example.
5) The following data relate to the number of family members in 30 families of a village.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 3 | 2 | 3 | 4 | 5 | 5 | 7 | 3 | 2 |
| 3 | 4 | 2 | 1 | 1 | 6 | 3 | 4 | 5 | 4 |
| 2 | 7 | 3 | 4 | 5 | 6 | 2 | 1 | 5 | 3 |

Classify the above data in the form of a discrete frequency distribution.

6) The profits (Rs. lakhs) of 50 companies are given below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 12 | 15 | 27 | 28 | 40 | 42 | 35 | 37 | 43 |
| 55 | 65 | 53 | 62 | 29 | 64 | 69 | 36 | 25 | 18 |
| 56 | 55 | 43 | 35 | 26 | 21 | 48 | 43 | 50 | 67 |
| 14 | 23 | 34 | 59 | 68 | 22. | 41 | 42 | 43 | 52 |
| 60 | 26 | 26 | 37 | 49 | 53 | 40 | 20 | 18 | 17 |

Classify the above data taking first class as 10-20 and form a frequency distribution.

7) The income (Rs.) of 24 employees of a company are given below:

| | | | | | |
|---|---|---|---|---|---|
| 1800 | 1250 | 1760 | 3500 | 6000 | 2500 |
| 2700 | 3600 | 3850 | 6600 | 3000 | 1500 |
| 4500 | 4400 | 3700 | 1900 | 1850 | 3750 |
| 6500 | 6800. | 5300 | 2700 | 4370 | 3300 |

ignou
THE PEOPLE'S
UNIVERSITY

Form a continuous frequency distribution after selecting a suitable class interval.

8) Draw a histogram and a frequency polygon from the following data:

| Marks | No. of students | Marks | No. of students |
|---|---|---|---|
| 0-20 | 8 | 60- 80 | 12 |
| 20-40 | 12 | 80-100 | 3 |
| 40-60 | 15 | | |

9) Go through the following data carefully and then construct a histogram.

| Income (Rs.) | No. of Persons | Income (Rs.) | No. of persons. |
|---|---|---|---|
| 500 1000 | 18 | 3000-4500 | |
| 1000-1500 | 20 | 4500-5000 | 12 |
| 1500-2500 | 30 | 5000-7000 | 5 |
| 2500-3000 | 25 | | |

10 The following data relating to sales of 100 companies is given below:

| Sales (Rs. lakhs) | No. of companies | Sales (Rs. lakhs) | No. of companies |
|---|---|---|---|
| 5-10 | 5 | 25-30 | 18 |
| 10-15 | 12 | 30-35 | 15 |
| 15-20 | 13 | 35-40 | 10 |
| 20-25 | 20 | 40-45 | 7 |

Draw less than and more than 0 gives. Determine the number of companies whose sales are (i) less than Rs.13 lakhs (ii) more than 36 lakhs and (iii) between Rs. 13 lakhs and Rs. 36 lakhs.

## 2.13 FURTHER READINGS

Clark, T.C. and E.W. Jordan. *Introduction to Business and Economic Statistics*, South-Western Publishing Co. Ohio, U.S.A.

Enns, P.G., *Business Statistics*, Richard D. Irwin Inc.: Homewood.

Gupta, S.P. and M.P. Gupta, *Business Statistics*, Sultan Chand & Sons.: New Delhi.

Levin, R.I., *Statistics for Management, Prentice*-Hall of India: New Delhi.

Moskowitz., H. and G.P. Wright. *Statistics for Management and Economics*, Charles. E. Merin Publishing Company: Ohio, U.S.A.

Edward Tufte *The Visual Display of Quantitative Information,* Graphic Press.

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 3 MEASURES OF CENTRAL TENDENCY

### Objectives

After going through this unit, you will learn:

- the concept and significance of measures of central tendency
- to compute various measures of central tendency, such as arithmetic mean, weighted arithmetic mean, median, mode, geometric mean and harmonic mean
- to compute several quantiles such as quartiles, deciles and percentiles
- the relationship among various averages.

### Structure

3.1 Introduction
3.2 Significance of Measures of Central Tendency
3.3 Properties of a Good Measure of Central Tendency
3.4 Arithmetic Mean
3.5 Mathematical Properties of Arithmetic Mean
3.6 Weighted Arithmetic Mean
3.7 Median
3.8 Mathematical Property of Median
3.9 Quantiles
3.10 Locating the Quantiles Graphically
3.11 Mode
3.12 Locating the Mode Graphically
3.13 Relationship among Mean, Median and Mode
3.14 Geometric Mean
3.15 Harmonic Mean
3.16 Summary
3.17 Key Words
3.18 Self-assessment Exercises
3.19 Further Readings

## 3.1 INTRODUCTION

With this unit, we begin our formal discussion of the statistical methods for summarising and describing numerical methods for summarising and describing numerical data. The objective here is to find one representative value which can be used to locate and summarise the entire set of varying values. This one value can be used to make many decisions concerning the entire set. We can define measures of central tendency (or location) to find some central value around which the data tend to cluster.

ignou THE PEOPLE'S UNIVERSITY

## 3.2 SIGNIFICANCE OF MEASURES OF CENTRAL TENDENCY

Measures of central tendency i.e. condensing the mass of data in one single value, enable us to get an idea of the entire data. For example, it is impossible to remember the individual incomes of millions of earning people of India. But if the average income is obtained, we get one single value that represents the entire population. Measures of central tendency also enable us to compare two or more sets of data to facilitate comparison. For example, the average sales figures of April may be compared with the sales figures of previous months.

## 3.3 PROPERTIES OF A GOOD MEASURE OF CENTRAL TENDENCY

A good measure of central tendency should possess, as far as possible, the following properties,

i) It should he easy to understand.

ii) It should he simple to compute.

iii) It should be based on all observations.

iv) It should be uniquely defined.

v) It should be capable of further algebraic treatment.

vi) It should not be unduly affected by extreme values.

Following are some of the important measures of central tendency which are commonly used in business and industry.

- Arithmetic Mean
- Weighted Arithmetic Mean
- Median
- Quantiles
- Mode
- Geometric Mean
- Harmonic Mean

## 3.4 ARITHMETIC MEAN

The arithmetic mean (or mean or average) is the most commonly used and readily understood measure of central tendency. In statistics, the term average refers to any of the measures of central tendency. The arithmetic mean is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations. Symbolically, it can be represented as:

$$\bar{X} = \frac{\sum x}{N}$$

ignou THE PEOPLE'S UNIVERSITY

where ∑x indicates the sum of the values of all the observations, and N is the total number of observations. For example, let us consider the monthly salary (Rs.) of 10 employees of a firm

2500, 2700, 2400, 2300, 2550, 2650, 2750, 2450, 2600, 2400

If we compute the arithmetic mean, then

$$\bar{X} = 2500 + 2700 + 2400 + 2300 + 2550 + 2650 + 2750 + 2450 + 2600 + 2400$$

$$= \frac{25300}{10} = Rs. 2530.$$

Therefore, the average monthly salary is Rs. 2530.

We have seen how to compute the arithmetic mean for ungrouped data. Now let us consider what modifications are necessary for grouped data. When the observations are classified into a frequency distribution, the midpoint of the class interval would be treated as the representative average value of that class. Therefore, for grouped data; the arithmetic mean is defined as

$$\bar{X} = \frac{\sum fx}{N}$$

Where X is midpoint of various classes, f is the frequency for corresponding class and N is the total frequency, i.e. N = $\sum f$.

This method is illustrated for the following data which relate to the monthly sales of 200 firms.

| Monthly Sales (Rs. Thousand) | No. of Firms | Monthly Sales (Rs. Thousand) | No. of Firms |
|---|---|---|---|
| 300-350 | 5 | 550-600 | 25 |
| 350-400 | 14 | 600-650 | 22 |
| 400-450 | 23 | 650-700 | 7 |
| 450-500 | 50 | 700-750 | 2 |
| 500-550 | 52 | | |

For computation of arithmetic mean, we need the following table:

| Monthly Sales (Rs. Thousand) | Mid point X | No. of firms f | fX |
|---|---|---|---|
| 300-350 | 325 | 5 | 1625 |
| 350-400 | 375 | 14 | 5250 |
| 400-450 | 425 | 23 | 9775 |
| 450-500 | 475 | 50 | 23750 |
| 500-550 | 525 | 52 | 27300 |
| 550-600 | 575 | 25 | 14375 |
| 600-650 | 625 | 22 | 13750 |
| 650-700 | 675 | 7 | 4725 |
| 700-750 | 725 | 2 | 1450 |
| | | N = 200 | ΣfX=102000 |

$$\bar{x} = \frac{\sum fx}{N} = \frac{102000}{200} = 510$$

Hence the average monthly sales are Rs. 510.

To simplify calculations, the following formula for arithmetic mean may be more convenient to use.

$$\bar{x} = A + \frac{\sum fd}{N} \times i$$

where A is an arbitrary point, $d = \frac{X-A}{i}$, and i = size of the equal class interval.

**REMARK:** A justification of this formula is as follows. When $d = \frac{X-A}{i}$, then X = A + i d Multiplying throughout by f, taking summation on both sides and. Dividing by N, we get

$$\bar{x} = A + \frac{\sum fd}{N} \times i$$

This formula makes the computations very simple and takes less time. To apply this formula, let us consider the same example discussed earlier and shown again in the following table.

| Monthly, Sales (Rs. Thousand) | Mid point | No. of Firms f | (X-525)/50 =d | fd |
|---|---|---|---|---|
| 300-350 | 325 | 5 | -4 | -20 |
| 350-400 | 375 | 14 | -3 | -42 |
| 400-450 | 425 | 23 | -2 | -46 |
| 450-500 | 475 | 50 | -1 | -50 |
| 500-550 | 525 | 52 | 0 | 0 |
| 550-600 | 575 | 25 | +1 | +25 |
| 600-650 | 625 | 22 | +2 | +44 |
| 650-700 | 675 | 7 | +3 | +21 |
| 700-750 | 725 | 2 | +4 | +8 |
| | | | N = 200 | ∑fd = –60 |

$$\bar{x} = A + \frac{\sum fd}{N} \times i = 525 - \frac{60}{200} \times 50$$

= 525 - 15 = 510 or Rs. 510

It may be observed that this formula is much faster than the previous one and the value of arithmetic mean remains the same.

## 3.5 MATHEMATICAL PROPERTIES OF ARITHMETIC MEAN

Because the arithmetic is defined operationally, it has several useful mathematical properties. Some of these are:

1) The sum of the deviations of the observations from the arithmetic mean is always zero. Symbolically, it is:

$$\sum(x - \bar{x}) = 0$$

It is because of this property that the mean is characterised as a point of balance, i.e, the sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

2) The sum of the squared deviations of the observations from the mean is minimum, i.e., the total of the squares of the deviations from any other value than the mean value will be greater than the total sum of squares of the deviations from mean. Symbolically,

   $\sum(x-\bar{x})^2$is a minimum.

3) The arithmetic means of several sets of data may be combined into a single arithmetic mean for the combined sets of data. For two sets of data, the combined arithmetic mean may be defined as

$$\bar{x}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Where $\bar{x}_{12}$ = combined mean of two sets of data.

$\bar{x}_{12}$ = arithmetic mean of the first set of data.

$\bar{x}_{12}$ = arithmetic mean of the second set of data.

$N_1$ = number of observations in the first set of data.

$N_2$ = number of observations in the second set of data.

If we have to combine three or more than three sets of data, then the same formula can be generalised as:

$$\bar{X}_{123.} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 + \cdots \ldots}{N_1 + N_2 + N_3 + \cdots \ldots}$$

The arithmetic mean has the great advantages of being easily computed and readily understood. It is due to the fact that it possesses almost all the properties of a good measure of central tendency. No other measure of central tendency possesses so many properties. However, the arithmetic mean has some disadvantages. The major disadvantage is that its value may be distorted by the presence of extreme values in a given set of data. A minor disadvantage is when it is used for open-end distribution since it is difficult to assign a midpoint value to the open-end class.

**Activity A**

The following data relate to the monthly earnings of 428 skilled employees in a big organisation. Compute the arithmetic mean and interpret this value.

| Monthly Earnings | No. of employees | Monthly Earnings | No. of employees |
|---|---|---|---|
| (Rs.) | (Rs.) | | |
| 1840-1900 | 1 | 2080-2140 | 126 |
| 1900-1960 | 3 | 2140-2200 | 90 |
| 1960-2020 | 46 | 220Q-2260 | 50 |
| 2020-2080 | 98 | 2260-2320 | 6 |
| | | 2320-2380 | 8 |

ignou THE PEOPLE'S UNIVERSITY

## 3.6 WEIGHTED ARITHMETIC MEAN

The arithmetic mean, as discussed earlier, gives equal importance (or weight) to each observation. In some cases, all observations do not have the same importance. When this is so, we compute weighted arithmetic mean. The weighted arithmetic mean can be defined as

$$\bar{X}_{\mathrm{w}} = \frac{\sum \mathrm{WX}}{\sum \mathrm{W}}$$

Where $\bar{X}_{\mathrm{w}}$ represents the weighted arithmetic mean,

W are the weights assigned to the variable X.

You are familiar with the use of weighted averages to combine several grades that are not equally important. For example, assume that the grades consist of one final examination and two mid term assignments. If each of the three grades are given a different weight, then the procedure is to multiply each grade (X) by its appropriate weight (W). If the final examination is 50 per cent of the grade and each mid term assignment is 25 per cent, then the weighted arithmetic mean is given as follows:

$$\bar{X}_{\mathrm{w}} = \frac{\sum \mathrm{WX}}{\sum \mathrm{W}} = \frac{\mathrm{W}_1\mathrm{X}_1 + \mathrm{W}_2\mathrm{X}_2 + \mathrm{W}_3\mathrm{X}_3}{\mathrm{W}_1 + \mathrm{W}_2 + \mathrm{W}_3}$$

$$= \frac{50\mathrm{X}_1 + 25\mathrm{X}_2 + 25\mathrm{X}_3}{50 + 25 + 25}$$

Suppose you got 80 in the final examination, 95 in the first mid term assignment, as 85 in the second mid term assignment then

$$\bar{X}_{\mathrm{w}} = \frac{50(80) + 25(95) + 25(85)}{100}$$

$$= \frac{4000 + 2375 + 2125}{100} = \frac{8500}{100} = 85$$

The following table shows this computation in a tabular form which is easy to employ for calculation of weighted arithmetic mean.

| | **Grade X** | **Weight W** | **WX** |
|---|---|---|---|
| Final Examination | 80 | 50 | 4000 |
| First assignment | 95 | 25 | 2375 |
| Second assignment | 85 | 25 | 2125 |
| | | $\sum \mathrm{W} = 100$ | $\sum \mathrm{WX} = 8500$ |

$$\bar{X}_{\mathrm{w}} = \frac{\sum \mathrm{WX}}{\sum \mathrm{W}} = \frac{8500}{100} = 85$$

The concept of weighted arithmetic mean is important because the computation is the same as used for averaging ratios and determining the mean of grouped data. Weighted mean is specially useful in problems relating to the construction of index numbers.

**Activity B**

A contractor employs three types of workers: male, female and children. He pays Rs. 40, Rs. 30, and Rs. 25 per day to a male, female and child worker respectively. Suppose he employs 20 males, 15 females, and 10 children. What is the average wage per day paid by the contractor? Would it make any difference in the answer if the number of males, females, and children employed are equal? Illustrate.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 3.7 MEDIAN

A second measure of central tendency is the median. Median is that value which divides the distribution into two equal parts. Fifty per cent of the observations in the distribution are above the value of median and other fifty per cent of the observations are below this value of median. The median is the value of the middle observation when the series is arranged in order of size or magnitude. If the number of observations is odd, then the median is equal to one of the original observations. If the number of observations is even, then the median is the arithmetic mean of the two middle observations. For example, if the income of seven persons in rupees is 1100, 1200, 1350, 1500, 1550, 1600, 1800, then the median income would be Rs. 1500. Suppose one more person joins and his income is Rs. 1850, then the median income of eight persons would be $\frac{1500+1550}{2} = 1525$ (since the number of observations is even, the median is the arithmetic mean of the 4th person and 5th person).

For grouped data, the following formula may be used to locate the value of median.

$$\text{Med.} = \mathrm{L} + \frac{\mathrm{N}/2 - \mathrm{pcf}}{\mathrm{f}} \times \mathrm{i}$$

where L is the lower limit of the median class, pcf is the preceding cumulative frequency to the median class, f is the frequency of the median class and i is the size of the median class.

As an illustration, consider the following data which relate to the age distribution of 1000 workers in an industrial establishment.

| Age (Years) | No. of workers | Age (Years) | No. of Workers |
|---|---|---|---|
| Below 25 | 120 | 40-45 | 150 |
| 25-30 | 125 | 45-50 | 140 |
| 30-35 | 180 | 50-55 | 100 |
| 35-40 | 160 | 55 and above | 25 |

Determine the median age.

The location of median value is facilitated by the use of a cumulative frequency distribution as shown below in the table.

| Age (Years) | No. of workers f | Cumulative frequency c.f |
|---|---|---|
| Below 25 | 120 | 120 |
| 25-30 | 125 | 245 |
| 30-35 | 180 | 425 |
| 35-40 | 160 | 585 |
| 40-45 | 150 | 735 |
| 45-50 | 140 | 875 |
| 50-55 | 100 | 975 |
| 55 and Above | 25 | 1000 |

N = 1000

Median class

Median = size of $\frac{N}{2}$ th observation = $\frac{1000}{2}$ = 500th observation which lies in the class 35 - 40.

Median = $L + \frac{N/2 - pcf}{f} \times i = 35 + \frac{500-425}{160} \times 5$

$= 35 + \frac{375}{160} = 35 + 2.34 = 37.34$ years.

Hence the median age is approximately 37 years. This value of median suggests that half of the workers are below the age of 37 years and other half of the workers are above the age of 37 years.

## 3.8 MATHEMATICAL PROPERTY OF MEDIAN

The important mathematical property of the median is that the sum of the absolute deviations about the median is a minimum. In symbols $\sum|X\text{-Med.}|$ is minimum.

Although the median is not as popular as the arithmetic mean, it does have the advantage of being both easy to determine and easy to explain.

As illustrated earlier, the median is affected by the number of observations rather than the values of the observations; hence it will be less distorted as a representative value than the arithmetic mean.

An additional advantage of the median is that it may be computed for an open-end distribution.

The major disadvantage of median is that further mathematical treatments cannot be done. However, since median is a positional average, its value is not determined by each and every observation.

**Activity C**

For the following data, compute the median and interpret this value.

| Monthly Rent (Rs.) | No. of Persons paying the rent | Monthly Rent (Rs.) | No. of Persons paying the rent |
|---|---|---|---|
| Below 1000 | 6 | 1800-2000 | 15 |
| 1000-1200 | 9 | 2000-2200 | 10 |
| 1200-1400 | 11 | 2200-2400 | 8 |
| 1400-1600 | 14 | 2400 and above | 7 |
| 1600-1800 | 20 | | |

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 3.9 QUANTILES

Quantiles are the related positional measures of central tendency. These are useful and frequently employed measures of non-central location. The most familiar quantiles are the **quartiles**, **deciles**, and **percentiles**.

**Quartiles:** Quartiles are those values which divide the total data into four equal parts. Since three points divide the distribution into four equal parts, we shall have three quartiles. Let us call them Q1, Q2, and Q3. The first quartile, Q1, is the value such that 25% of the observations are smaller and 75% of the observations are larger. The second quartile, Q2, is the median, i.e., 50% of the observations are smaller and 50% are larger. The third quartile, Q3, is the value such that 75% of the observations are smaller and 25% of the observations are larger.

For grouped data, the following formulas are used for quartiles.

$$Q_j = L + \frac{jN/4 - pcf}{f} \times i \qquad \text{for } j = 1,2,3$$

where L is lower limit of the quartile class, pcf is the preceding cumulative frequency to the quartile class, f is the frequency of the quartile class, and i is the size of the quartile class.

**Deciles:** Deciles are those values which divide the total data into ten equal parts. Since nine points divide the distribution into ten equal parts, we shall have nine deciles denoted by $D_1$, $D_2$, , $D_9$,

For grouped data, the following formulas are used for deciles:

$$D_k = L + \frac{KN/10 - pcf}{f} \times i \qquad k = 1,2,\ldots\ldots,9$$

where the symbols have usual meaning and interpretation.

**Percentiles:** Percentiles are those values which divide the total data into hundred equal parts. Since ninety nine points divide the distribution into hundred equal parts, we shall have ninety nine percentiles denoted by

$P_1, P_2, P_3, \dots \dots \dots \dots \dots \dots, P_{99}$

For grouped data, the following formulas are used for percentiles.

$P_l = L + \frac{lN/100 - pcf}{f} \times i$ for $l = 1, 2, \dots, 99$

To illustrate the computations of quartiles, deciles and percentiles, consider the following grouped data which relate to the profits of 100 companies during the year 1987-88.

| **Profits (Rs. lakhs)** | **No. of companies** | **Profits (Rs. lakhs)** | **No. of companies** |
|---|---|---|---|
| 20-30 | 4 | 60-70 | 15 |
| 30-40 | 8 | 70-80 | 10 |
| 40-50 | 18 | 80-90 | 8 |
| 50-60 | 30 | 90-100 | 7 |

Calculate $Q_1$, $Q_2$, (median), $D_6$, and $P_{90}$, from the given data and interpret these values.

To compute $Q_1$, $Q_2$, $D_6$, and $P_{90}$, we need the following table:

| **Profits (Rs. lakhs)** | **No. of companies ($f$)** | **c.f** |
|---|---|---|
| 20-30 | 4 | 4 |
| 30-40 | 8 | 12 |
| 40-50 | 18 | 30 ⟵ |
| 50-60 | 30 | 60 |
| 60-70 | 15 | 75 |
| 70-80 | 10 | 85 |
| 80-90 | 8 | 93 |
| 90-100 | 7 | 100 |

$Q_1$ = Size of N/4th observation $= \frac{100}{4} = 25^{\text{th}}$ observation, which lies in the class 40 — 50

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 40 + \frac{25 - 12}{18} \times 10 = 40 + 7.22 = 47.22$$

This value of $Q_1$ suggests that 25% of the companies earn an annual profit of Rs. 47.22 lakh or less.

Median or $Q_2$ = Size of $\frac{N}{2}$ th observation $= \frac{100}{2}$ = 50th observation which lies in the class 50 — 60.

$$Q_2 = L + \frac{N/2 - pcf}{f} \times i = 50 + \frac{50 - 30}{30} \times 10 = 50 + 6.67 = 56.67$$

This value of $Q_2$, (or median) suggests that-50% of the companies earn an annual profit of Rs. 56.67 lakh or less and the remaining 50% of the companies earn an annual profits of Rs. 56.67 lakh or more.

$D_6$ = Size of $\frac{6N}{10}$ th observation = $\frac{6\times100}{10}$ = 60th observation, which lies in the class 50 — 60.

$$D_6 = L + \frac{6N/10 - pcf}{f} \times i = 50 + \frac{60 - 30}{30} \times 10 = 50 + 10 = 60$$

Thus 60% of the companies earn an annual profit of Rs. 60 lakh or less and 40% of the companies earn Rs. 60 lakh or more.

$P_{90}$ = size of $\frac{90N}{100}$th observation = $\frac{90\times100}{100}$ = 90th observation, which lies in the class 80-90.

$$P_{90} = L + \frac{90N/100 - pcf}{f} \times i = 80 + \frac{90 - 85}{10} \times 10 = 80 + 5 = 85$$

This value of 90th percentile suggests that 90% of the companies earn an annual profit of Rs. 85 lakh or less and 20% of the companies earn more than Rs. 85 lakh or more.

## 3.10 LOCATING THE QUANTILES GRAPHICALLY

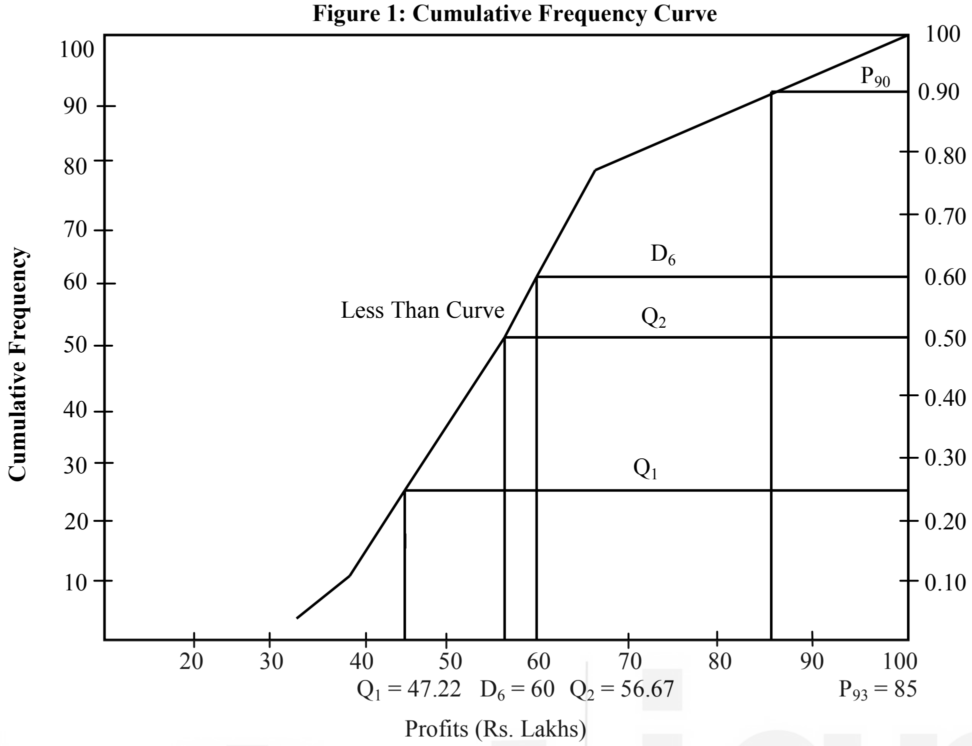
To locate the median graphically, draw less than cumulative frequency curve (less than ogive). Take the variable on the X-axis and frequency on the Y-axis. Determine the median value by locating N/2th observation on the Y-axis. Draw a horizontal line from this on the cumulative frequency curve and from where it meets the curve, draw a perpendicular on the X-axis. The point where it meets the X-axis is the value of median.

Similarly we can locate graphically the other quantiles such as quartiles, deciles and percentiles.

For the data of previous illustration, locate graphically the values of $Q_1$, $Q_2$, $D_{60}$, and $Q_{90}$.

The first step is to make a less than cumulative frequency curve as shown in figure I.

Figure 1: Cumulative Frequency Curve



To determine different quantiles graphically, horizontal lines are drawn from the cumulative relative frequency values. For example if we want to determine the value of median (or $Q_2$), a horizontal line can be drawn from the cumulative frequency value of 0.50 to the less than curve and then extending the vertical line to the horizontal axis. Ina similar way, other values can be determined as shown in the graph. From the graph, we observe

$$Q_1 = 47.22, Q_2 = 57.67, D_2 = 60.0, P_{90} = 85$$

It may be noted that these graphical values of quantiles are the same as obtained by the formulas.

**Activity D**

Given below is the wage distribution of 100 workers in a factory:

| Wages (Rs.) | No. of workers | Wages (Rs.) | No. of workers |
|---|---|---|---|
| Below 1000 | 3 | 1800-2000 | 10 |
| 1000-1200 | 5 | 2000-2200 | 8 |
| 1200-1400 | 12 | 2200-2400 | 5 |
| 1400-1600 | 23 | 2400 and above | 3 |
| 1600-1800 | 31 | | |

Draw a less than cumulative frequency curve (ogive) and use it to determine graphically the values of $Q_2$, $Q_3$, $D_{60}$, and $P_{80}$. Also verify your result by the corresponding mathematical formula.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 3.11 MODE

The mode is the typical or commonly observed value in a set of data. It is defined as the value which occurs most often or with the greatest frequency. The dictionary meaning of the term mode is most usual. For example, in the series of numbers 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, the mode is 8 because it occurs the maximum number of times.

The calculations are different for the grouped data, where the modal class is defined as the class with the maximum frequency. The following formula is used for calculating the mode.

$$\text{Mode} = L + \frac{d_1}{d_1+d_2} \times i$$

where L is lower limit of the modal class, $d_1$ is the difference between the frequency of the modal class and the frequency of the preceding class, $d_2$ is the difference between the frequency of the modal class and the frequency of the succeeding class, i is the size of the modal class. To illustrate the computation of mode, let us consider the following data.

| Daily Sales (Rs. thousand) | No. of firms | Daily Sales (Rs. thousand) | No. of firms |
|---|---|---|---|
| 20-30 | 15 | 60-70 | 35 |
| 30-40 | 23 | 70-80 | 25 |
| 40-50 | 27 | 80-90 | 5 |
| 50-60 | 20 | | |

Since the maximum frequency 35 is in the class 60-70, therefore 60-70 is the modal class. Applying the formula, we get

$$\text{Mode} = L + \frac{d_1}{d_1+d_2} \times i = 60 + \frac{35-20}{(35-20)+(35-25)} \times 10$$

$$= 60 + \frac{150}{25}$$

$$= 60 + 6 = \text{Rs.}66.$$

Hence modal daily sales are Rs. 66.

## 3.12 LOCATING THE MODE GRAPHICALLY

In a grouped data, the value of mode can also be determined graphically. In graphical method, the first step is to construct histogram for the given data. The next step is to draw two straight lines diagonally on the inside of the modal class bars, starting from each upper corner of the bar to the upper corner of the adjacent bar. The last step is to draw a perpendicular line from the intersection of the two diagonal lines to the X-axis which gives us the modal value.

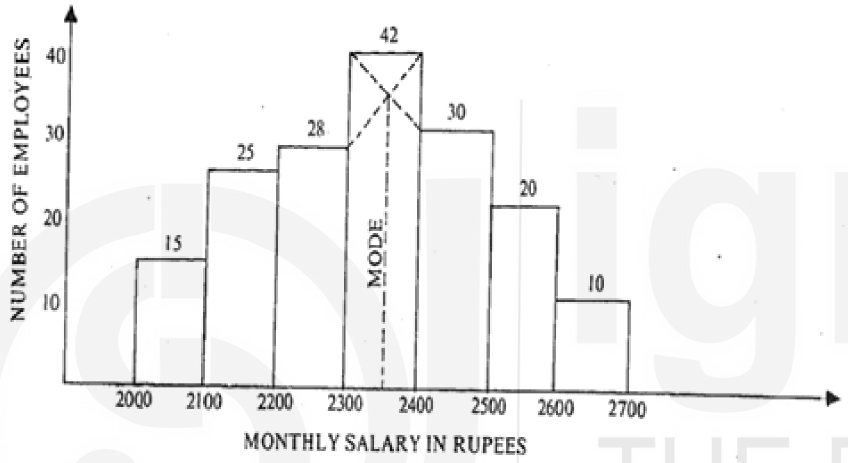
Consider the following data to locate the value of mode graphically.

| Monthly salary (Rs.) | No. of employees | Monthly salary (Rs.) | No. of employees |
|---|---|---|---|
| 2000-2100 | 15 | 2400-2500 | 30 |
| 2100-2200 | 25 | 2500-2600 | 20 |
| 2200-2300 | 28 | 2600-2700 | 10 |
| 2300-2400 | 42 | | |

First draw the histogram as shown below in figure II.

Figure II: Histogram of Monthly Salaries

**Figure II: Histogram of Monthly Salaries**



The two straight lines are drawn diagonally in the inside of the modal class bars and then finally a vertical line from the intersection of the two diagonal lines is drawn on the X-axis. Thus the modal value is approximately Rs. 2353. It may be noted that the value of mode would be approximately the same if we use the algebric method.

The chief advantage of the mode is that it is, by definition, the most representative value of the distribution. For example, when we talk of modal size of shoe or garment, we have this average in mind. Like median, the value of mode is not affected by extreme values and its value can be determined in open-end distributions.

The main disadvantage of the mode is its indeterminate value, i.e., we cannot calculate its value precisely in a grouped data, but merely estimate it. When a given set of data have two or more than two values as maximum frequency, it is a case of bimodal or multimodal distribution and the value of mode is not unique. The mode has no useful mathematical properties. Hence, in actual practice the mode is more important as a conceptual idea than as a working average.

**Activity E**

Compute the value of mode from the grouped data given below. Also check this value of mode graphically.

ignou
THE PEOPLE'S
UNIVERSITY

| Monthly stipend (Rs.) | No. of management trainees | Monthly stipend (Rs.) | No. of trainees |
|---|---|---|---|
| 2500-2700 | 25 | 3300-3500 | 20 |
| 2700-2900 | 35 | 3500-3700 | 15 |
| 2900-3100 | 60 | 3700-3900 | 5 |
| 3100-3300 | 40 | | |

…………………………………………………………………………………..

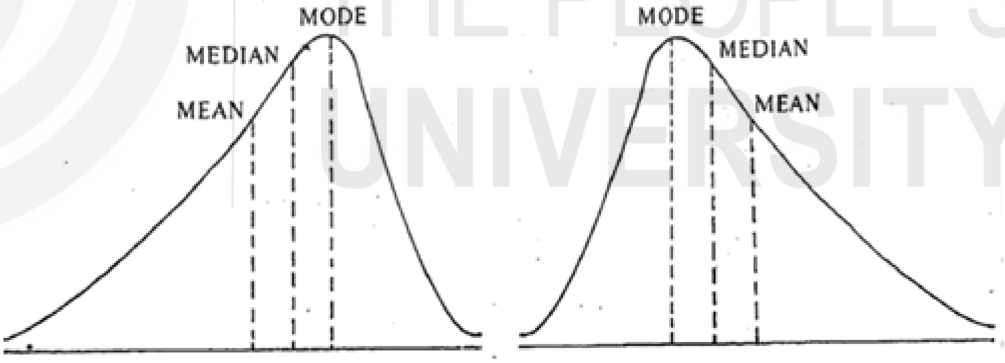
…………………………………………………………………………………..

…………………………………………………………………………………..

…………………………………………………………………………………..

## 3.13 RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

A distribution in which mean, median and mode coincide is known as a symmetrical (bell shaped) distribution. If a distribution is skewed (that is, not symmetrical) then mean, median, and mode are not equal. In a moderately skewed distribution, a very interesting relationship exists among mean, median and mode. In such type of distributions, it can be proved that the distance between mean and median is approximately one third of the distance between the mean and mode. This is shown below for two types of such distributions.



This relationship can be expressed as follows:

Mean - Median = 1/3 (Mean - Mode)

or Mode = 3 Median - 2 Mean

Similarly, we can express the approximate relationship for median in terms of mean and mode. Also this can be expressed for mean in terms of median and mode. Thus, if we know any of the two values of the averages, the third value of the average can be determined from this approximate relationship.

For example, consider a moderately skewed distribution in which mean and median is 35.4 and 34.3 respectively. Calculate the value of mode.

To compute the value of mode, we use the approximate relationship

Mode = 3 Median - 2 Mean

= 3 (34.3) - 2 (35.4)

= 102.9-70.8 = 32.1

Therefore the value of mode is 32.1.

## 3.14 GEOMETRIC MEAN

The geometric mean like the arithmetic mean, is a calculated average. The geometric mean, GM, of a series of numbers, $X_1$ $X_2$, .... $X_n$, is defined as

$$GM = \sqrt[N]{X_1.X_2.X_3\ldots\ldots\ldots X_N}$$

or the $N^{th}$ root of the product of N observations.

When the number of observations is three or more, the task of computation becomes quite tedious. Therefore a transformation into logarithms is useful to simplify calculations. If we take logarithms of both sides, then the formula for GM becomes

$$\text{Log GM} = \frac{1}{N}(\log X_1 + \log X_2 + \ldots\ldots + \log X_N)$$

and therefore, $GM = \text{Antilog}\left(\frac{\sum \log X}{N}\right)$

For the grouped data, the geometric mean is calculated with the following formula

$$GM = \text{Antilog}\left[\frac{\sum f(\log X)}{N}\right]$$

Where the notation has the usual meaning.

Geometric mean is specially useful in the construction of index numbers. It is an average most suitable when large weights have to be given to small values of observations and small weights to do large values of observations. This average is also useful in measuring the growth of population.

The following data illustrates the use and the computations involved in geometric mean.

A machine was purchased for Rs. 50,000 in 1984. Depreciation on the diminishing balance was charged @ 40% in the first year, 25% in the second year and 15% per annum during the next three years. What is the average depreciation charged during the whole period?

Since we are interested in finding the average rate of depreciation, geometric mean will be the most appropriate average.

| Year | Diminishing value (for a value of Rs. 100) X | Log X |
|---|---|---|
| 1984 | 100 - 40 = 60 | 1.77815 |
| 1985 | 100 - 25 = 75 | 1.87506 |
| 1986 | 100-15 = 85 | 1.92941 |
| 1987 | 100- 15 = 85 | 1.92941 |
| 1988 | 100-15 = 85 | 1.92941 |
| | | $\sum \log X = 9.44144$ |

$$GM = \text{Antilog}\left(\frac{\sum \log X}{N}\right)$$

$$= \text{Antilog}\left(\frac{9.44144}{5}\right) = \text{Antilog } 1.8883 = 77.32$$

The diminishing value being Rs. 77.32, the depreciation will be 100-77.32 = 22.68%. The geometric mean is very useful in averaging ratios and percentages. It also helps in determining the rates of increase and decrease. It is also capable of further algebraic treatment, so that a combined geometric mean can easily be computed. However, compared to arithmetic mean, the geometric mean is more difficult to compute and interpret. Further, geometric mean cannot be computed if any observation has either a value zero or negative:

**Activity F**

Find the geometric mean for the following data:

| Class interval | Frequency | Class interval | Frequency |
|---|---|---|---|
| 4.5-5.5 | 8 | 8.5- 9.5 | 25 |
| 5.5-6.5 | 10 | 9.5 - 10.5 | 18 |
| 6.5-7.5 | 12 | 10.5-11.5 | 7 |
| 7.5 - 8.5 | 15 | 11.5-12.5 | 5 |

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 3.15 HARMONIC MEAN

The harmonic mean is a measure of central tendency for data expressed as rates such as kilometers per hour, tonnes per day, kilometers per litre etc. The harmonic mean is defined as the **reciprocal of the arithmetic mean of the reciprocal of the individual observations**. If $X_1$, $X_2$, ……….. $X_N$ are N observations, then harmonic mean can be represented by the following formula.

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots \ldots .. + \frac{1}{X_N}} = \frac{N}{\sum\left(\frac{1}{X}\right)}$$

For example, the harmonic mean of 2, 3, 4 is

$$\mathrm{HM} = \frac{3}{\frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = \frac{3}{13/12} = \frac{36}{13} = 2.77$$

For grouped data, the formula becomes

$$\mathrm{HM} = \frac{\mathrm{N}}{\sum\left(\frac{\mathrm{f}}{\mathrm{X}}\right)}$$

The harmonic mean is useful for computing the average rate of increase of profits, or average speed at which a journey has been performed, or the average price at which an article has been sold. Otherwise its field of application is really restricted.

To explain the computational procedure, let us consider the following example.

In a factory, a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes, and by E in 12 minutes. Find the average number of units of work completed per minute.

The calculations for computing harmonic mean are given below:

| X | 1/X |
|---|---|
| 4 | 0.250 |
| 5 | 0.200 |
| 6 | 0.167 |
| 10 | 0.100 |
| 12 | 0.083 |
|  | $\sum 1/X = 0.8$ |

Hence the average number of units computed per minute is 5/0.8 = 6.25.

The harmonic mean like arithmetic mean and geometric mean is computed from each and every observation. It is specially useful for averaging rates.

However, harmonic mean cannot be computed when one or more observations have zero value or when there are both positive or negative observations. In dealing with business problems, harmonic mean is rarely used.

**Activity G**

In a factory, four workers are assigned to complete an order received for dispatching 1400 boxes of a particular commodity. Worker-A takes 4 minutes per box, B takes 6 minutes per box, C takes 10 minutes per box, D takes 15 minutes per box. Find the average minutes taken per box by the group of workers.

…………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 3.16 SUMMARY

Measures of central tendency give one of the very important characteristics of data. Any one of the various measures of central tendency may be chosen as the most representative or typical measure. The arithmetic mean is widely used and understood as a measure of central tendency. The concepts of weighted arithmetic mean, geometric mean, and harmonic mean are useful for specified type of applications. The median is generally a more representative measure for open-end distribution and highly skewed distribution. The mode should be used when the most demanded or customary value is needed.

## 3.17 KEY WORDS

**Arithmetic Mean** is equal to the sum of the values divided by the number of values.

**Geometric Mean** of N observations is the Nth root of the product of the given value observations.

**Harmonic Mean** of N observations is the reciprocal of the arithmetic mean of the reciprocals of the given values of N observations.

**Median** is that value of the variable which divides the distribution into two equal parts.

**Mode** is that value of the variable which occurs the maximum number of times.

**Quantiles** are those values which divide the distribution into a fixed number of equal parts, eg., quartiles divide distribution into four equal parts.

## 3.18 SELF-ASSESSMENT EXERCISES

1) List the various measures of central tendency studied in this unit and explain the difference between them.
2) Discuss the mathematical properties of arithmetic mean and median.
3) Review for each of the measure of central tendency, their advantages and disadvantages.
4) Explain how you will decide which average to use in a particular problem.
5) What are quantiles? Explain and illustrate the concepts of quartiles, deciles and percentiles.

6) Following is the cumulative frequency distribution of preferred length of study table obtained from the preferency study of 50 students.

| Length | No. of students | Length | No. of students |
|---|---|---|---|
| more than 50 cms | 50 | more than 90 cms | 25 |
| more than 60 cms | 46 | more than 100 cms | 18 |
| more than 70 cms | 40 | more than 110 cms | 7. |
| more than 80 cms | 32 | | |

A manufacturer has to take decision on the length of study-table to manufacture. What length would you recommend and why?

7) A three month study of the phone calls received by Small Company yielded the following information.

| Number of calls per day | No. of days | Number of calls per day | No. days |
|---|---|---|---|
| 100 - 200 | 3 | 600- 700 | 10 |
| 200-300 | 7 | 700- 800 | 9 |
| 300-400 | 11 | 800-900 | 8 |
| 400-500 | 13 | 900- 1000 | 4 |
| 500- 600 | 27 | | |

Compute the arithmetic mean, median and mode.

From the following distribution of travel time of 213 days to work of a firm's find the modal travel time.

| Travel time (in minutes) | No. of Days | Travel time (in minutes) | No. of days |
|---|---|---|---|
| Less than 80 | 213 | Less than 40 | 85 |
| Less than 70 | 210 | Less than 30 | 50 |
| Less than 60 | 195 | Less than 20 | 13 |
| Less than 50 | 156 | Less than 10 | 2 |

8) The mean monthly salary paid to all employees in a company is Rs. 1600. The mean monthly salaries paid to technical employees are Rs. 1800 and Rs. 1200 respectively. Determine the percentage of technical and non-technical employees of the company.

9) The following distribution is with regard to weight (in grams) of apples of a given variety. If an apple of less than 122 grams is to be considered unsuitable for export, what is the percentage of total apples suitable for the export?

| Weight (in grams) | No. of apples | Weight (in grams) | No. of apples |
|---|---|---|---|
| 100-110 | 10 | 140-150 | 35 |
| 110-120 | 20 | 150-160 | 15 |
| 120-130 | 40 | 160-170 | 5 |
| 130-140 | | | |

ignou
THE PEOPLE'S
UNIVERSITY

Draw an ogive of more than one type and deduce how many apples will be more than 122 grams.

10) The geometric mean of 10 observations on a certain variable was calculated to be 16.2. It was later discovered that one of the observations was wrongly recorded as 10.9 when in fact it was 21.9. Apply appropriate correction and calculate the correct geometric mean

11) An incomplete distribution of daily sales (Rs. thousand) is given below. The data relate to 229 days.

| **Daily sales (Rs. thousand)** | **No. of days** | **Daily sales (Rs. thousand)** | **No. of days** |
|---|---|---|---|
| 10-20 | 12 | 50-60 | ? |
| 20-30 | 30 | 60-70 | 25 |
| 30-40 | ? | 70-80 | 18 |
| 40 -50 | | | |

You are told that the median value is 46. Using the median formula, fill up the missing frequencies and calculate the arithmetic mean of the completed data.

12) The following table shows the income distribution of a company.

| **Income (Rs.)** | **No. of employees** | **Income (Rs.)** | **No. of employees** |
|---|---|---|---|
| 1200-1400 | 8 | 2200-2400 | 35 |
| 1400-1600 | 12 | 2400-2600 | 18 |
| 1600-1800 | 20 | 2600-2800 | 7 |
| 1800-2000 | 30 | 2800-3000 | 6 |
| 2000-2200 | 40 | 3000-3200 | 4 |

**Determine** (i) the mean income (ii) the median income (iii) the mean (iv) the income limits for the middle 50% of the employees (v) D7, the seventh docile, and (vi) P80, the eightieth percentile.

## 3.19 FURTHER READINGS

Clark, T.C. and E. W. Jordan. Introduction to Business and Economic Statistics, South-Western Publishing Co.

Enns, P.G., Business Statistics. Richard D. Irwin: Homewood.

Gupta, S.P. and M.P. Gupta, Business Statistics, Sultan Chand & Sons: New Delhi.

Moskowitz, H. and G.P. Wright, Statistics for Management and Economics, Charles E. Merin Publishing Company:

B. Bowerman and Richad O' Cennell, Business statistics in Practice, McGraw Hill.

# UNIT 4 MEASURES OF VARIATION AND SKEWNESS

## Objectives

After going through this unit, you will learn:

- the concept and significance of measuring variability
- the concept of absolute and relative variation
- the computation of several measures of variation, such as the range, quartile deviation, average deviation and standard deviation and also their coefficients
- the concept of skewness and its importance
- the computation of coefficient of skewness.

## Structure

4.1 Introduction
4.2 Significance of Measuring Variation
4.3 Properties of a Good Measure of Variation
4.4 Absolute and Relative Measures of Variation
4.5 Range
4.6 Quartile Deviation
4.7 Average Deviation
4.8 Standard Deviation
4.9 Coefficient of Variation
4.10 Skewness
4.11 Relative Skewness
4.12 Summary
4.13 Key Words
4.14 Self-assessment Exercises
4.15 Further Readings

ignou
THE PEOPLE'S
UNIVERSITY

## 4.1 INTRODUCTION

In the previous unit, we were concerned with various measures that are used to provide a single representative value of a given set of data. This single value alone cannot adequately describe a set of data. Therefore, in this unit, we shall study two more important characteristics of a distribution. First we shall discuss the concept of variation and later the concept of skewness.

A measure of variation (or dispersion) describes the spread or scattering of the individual values around the central value. To illustrate the concept of variation, let us consider the data given below:

| Firm A<br>Daily Sales (Rs.) | Firm B<br>Daily Sales (Rs.) | Firm C<br>Daily Sales (Rs.) |
|---|---|---|
| 5000 | 5050 | 4900 |
| 5000 | 5025 | 3100 |
| 5000 | 4950 | 2200 |
| 5000 | 4835 | 1800 |
| 5000 | 5140 | 13000 |
| $\overline{X}_A = 5000$ | $\overline{X}_B = 5000$ | $\overline{X}_c = 5000$ |

Since the average sales for firms A, B and C is the same, we are likely to conclude that the distribution pattern of the sales is similar. It may be observed that in Firm A, daily sales are the same irrespective of the day, whereas there is less amount of variation in the daily sales for firm 13 and greater amount of variation in the daily sales for firm C. Therefore, different sets of data may have the same measure central tendency but differ greatly in terms of variation.

## 4.2 SIGNIFICANCE OF MEASURING VARIATION

Measuring variation is significant for some of the following purposes.

i) Measuring variability determines the reliability of an average by pointing out as to how far an average is representative of the entire. data.

ii) Another purpose of measuring variability is to determine the nature and cause variation in order to control the variation itself.

iii) Measures of variation enable comparisons of two or more distributions with regard to their variability.

iv) Measuring variability is of great importance to advanced statistical analysis. For example, sampling or statistical inference is essentially a problem in measuring variability.

## 4.3 PROPERTIES OF A GOOD MEASURE OF VARIATION

A good measure of variation should possess, as far as possible, the same properties as those of a good measure of central tendency.

Following are some of the well known measures of variation which provide a numerical index of the variability of the given data:

i) Range

ii) Average or Mean Deviation

iii) Quartile Deviation or Semi-Interquartile Range

iv) Standard Deviation

## 4.4 ABSOLUTE AND RELATIVE MEASURES OF VARIATION

Measures of variation may be either absolute or relative. Measures of absolute variation are expressed in terms of the original data. In case the two sets of data are expressed in different units of measurement, then the absolute measures of variation are not comparable. In such cases, measures of relative variation should be used. The other type of comparison for which measures of relative variation are used involves the comparison between two sets of data having the same unit of measurement but with different means. We shall now consider in turn each of the four measures of variation.

## 4.5 RANGE

The range is defined as the difference between the highest (numerically largest) value and the lowest (numerically smallest) value in a set of data. In symbols, this may be indicated as:

$R = H - L$,

where R = Range; H = Highest Value; L = Lowest Value

As an illustration, consider the daily sales data for the three firms as given earlier.

For firm A, $R = H - L = 5000 - 5000 = 0$

For firm B, $R = H - L = 5140 - 4835 = 305$

For firm C, $R = H - L = 13000 - 1800 = 11200$

The interpretation for the value of range is very simple.

In this example, the variation is nil in case of daily sales for firm A, the variation is small in case of firm B and variation is very large in case of firm C.

The range is very easy to calculate and it gives us some idea about the variability of the data. However, the range is a crude measure of variation, since it uses only two extreme values.

The concept of range is extensively used in statistical quality control. Range is helpful in studying the variations in the prices of shares and debentures and other commodities that are very sensitive to price changes from one period to another. For meteorological departments, the range is a good indicator for weather forecast.

For grouped data, the range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula

Coefficient of range $= \frac{H-L}{H+L}$

**Activity A**

Following are the prices of shares of a company from Monday to Friday:

| Day | : | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|---|
| Price | : | 670 | 678 | 750 | 705 | 720 |

Compute the value of range and interpret the value.

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

**Activity B**

Calculate the coefficient of range from the following data:

| Sales (Rs. lakhs) | No. of Days | Sales (Rs. lakhs) | No. of days |
|---|---|---|---|
| 30-40 | 12 | 60-70 | 19 |
| 40-50 | 18 | 70-80 | 13 |
| 50-60 | 20 | 80-90 | 8 |

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

## 4.6 QUARTILE DEVIATION

The quartile deviation, also known as semi-interquartile range, is computed by taking the average of the difference between the third quartile and the first quartile. In symbols, this can be written as:

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

where $Q_1$ = first quartile, and $Q_3$ = third quartile.

The following illustration would clarify the procedure involved. For the data given below, compute the quartile deviation.

| **Monthly Wages (Rs.)** | **No. of workers** | **Monthly Wages (Rs.)** | **No. of Workers** |
|---|---|---|---|
| Below 850 | 12 | 1000-1050 | 62 |
| 850-900 | 16 | 1050-1100 | 75 |
| 900-950 | 39 | 1100-1150 | 30 |
| 950-1000 | 56 | 1150 and above | 10 |

To compute quartile deviation, we need the values of the first quartile and the third quartile which can be obtained from the following table:

| **Monthly Wages (Rs.)** | **No. of workers f** | **C.F.** |
|---|---|---|
| Below 850 | 12 | 12 |
| 850-900 | 16 | 28 |
| 900-950 | 39 | 67 |
| 950 -1000 | 56 | 123 |
| 1000-1050 | 62 | 185 |
| 1050-1100 | 75 | 260 |
| 1100-1150 | 30 | 290 |
| 1150 and above | 10 | 300 |

$Q_1$ = Size of $\frac{N}{4}$th observation $= \frac{300}{4}$ = 75th observation which lies in the class $950 - 1000$

$$Q_1 = L + \frac{N/4 - \text{pcf}}{f} \times i = 950 + \frac{75 - 67}{56} \times 50$$

$$= 950 + \frac{50}{7} = 950 + 7.14 = 957.14$$

$Q_3$ = Size of $\frac{3N}{4}$th observation $= \frac{3\times300}{4}$ = $225^{\text{th}}$ observation which lies in the class $1050 - 1100$

$$Q_3 = L + \frac{3N/4 - \text{pcf}}{f} \times i = 1050 + \frac{225 - 185}{75} \times 50$$

$$= 1050 + \frac{2000}{75} = 1050 + 26.67 = 1076.67$$

$$\text{Q.D.} = \frac{1076.67 - 957.14}{2} = \frac{119.53}{2} = 59.765$$

The relative measure corresponding to quartile deviation, called the coefficient of quartile deviation, is calculated as given below:

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

The quartile deviation is superior to the range as it is not based on two extreme values but rather on middle 50% observations. Another advantage of quartile deviation is that it is the only measure of variability which can be used for open-end distribution.

The disadvantage of quartile deviation is that it ignores the first and the last 25% observations.

**Activity C**

A survey of domestic consumption of electricity gave the following distribution of the units consumed. Compute the quartile deviation and its coefficient.

| Number of units | Number of consumers | Number of units | Number of consumers |
|---|---|---|---|
| Below 200 | 9 | 800-1000 | 45 |
| 200-400 | 18 | 1000-1200 | 38 |
| 400-600 | 27 | 1200-1400 | 20 |
| 600-800 | 32 | 1400 & above | 11 |

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

……………………………………………………………………………..

## 4.7 AVERAGE DEVIATION

The measure of average (or mean) deviation is an improvement over the previous two measures in that it considers all observations in the given set of data. This measure is computed as the mean of deviations from the mean or the median. All the deviations are treated as positive regardless of sign. In symbols, this can be represented by:

$$\text{A.D.} = \frac{\sum|X-\bar{X}|}{N} \text{ or } \frac{\sum|X-\text{ Median }|}{N}$$

Theoretically speaking, there is an advantage in taking the deviations from median because the sum of the absolute deviations (i.e. ignoring ± signs) from median is minimum. In actual practice, however, arithmetic mean is more popularly used in computation of average deviation.

For grouped data, the formula to be used is given as:

$$\text{A.D.} = \frac{\sum|X-\bar{X}|}{N}$$

As an illustration, consider the following grouped data which relate to the sales of 100 companies.

| Sales (Rs. thousand) | No. of days | Sales (Rs. thousand) | No. of days |
|---|---|---|---|
| 40-50 | 10 | 70-80 | 30 |
| 50-60 | 15 | 80-90 | 12 |
| 60-70 | 25 | 90-100 | 8 |

To compute average deviation, we construct the following table:

| Sales (Rs. thousand) | X m.p | No. of days | fX | $\lvert X-\bar{X}\rvert$ | $f\lvert X-\bar{X}\rvert$ |
|---|---|---|---|---|---|
| 40-50 | 45 | 5 | 225 | 26 | 130 |
| 50-60 | 55 | 15 | 825 | 16 | 240 |
| 60-70 | 65 | 25 | 1625 | 6 | 150 |

ignou THE PEOPLE'S UNIVERSITY

| | | | | | |
|---|---|---|---|---|---|
| 70-80 | 75 | 30 | 2250 | 4 | 120 |
| 80-90 | 85 | 20 | 1700 | 14 | 280 |
| 90-100 | 95 | 5 | 475 | 24 | 120 |
| | | $N = 100$ | $\sum fX = 7100$ | | $\Sigma f\lvert X-\bar{X}\rvert = 1040$ |

$$\bar{X} = \frac{\sum fX}{N} = \frac{7100}{100} = 71$$

$$A.D = \frac{\Sigma f\lvert X-\bar{X}\rvert}{N} = \frac{1040}{100} = 10.4$$

The relative measure corresponding to the average deviation, called the coefficient of average deviation, is obtained by dividing average deviation by the particular average used in computing the average deviation. Thus, if average deviation has been computed from median, the coefficient of average deviation shall be obtained by dividing the average deviation by the median.

$$\text{Coefficient of A.D.} = \frac{\text{A.D.}}{\text{Median}} \text{ or } \frac{\text{A.D.}}{\text{Mean}}$$

Although the average deviation is a good measure of variability, its use is limited. If one desires only to measure and compare variability among several sets of data, the average deviation may be used.

The major disadvantage of the average deviation is its lack of mathematical properties. This is more true because non-use of signs in its calculations makes it algebraically inconsistent.

**Activity D**

Calculate the average deviation and coefficient of the average deviation from the following data.

| Sales (Rs. thousand) | No. of days | Sales (Rs. thousand) | No. of days |
|---|---|---|---|
| Less than 20 | 3 | Less than 50 | 23 |
| Less than 30 | 9 | Less than 60 | 25 |
| Less than 40 | 20 | | |

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

ignou THE PEOPLE'S UNIVERSITY

## 4.8 STANDARD DEVIATION

The standard deviation is the most widely used and important measure of variation. In computing the average deviation, the signs are ignored. The standard deviation overcomes this problem by squaring the deviations, which makes them all positive. The standard deviation, also known as root mean square deviation, is generally denoted by the lower case Greek letter a (read as sigma). In symbols, this can be expressed as

$$\sigma = \sqrt{\frac{\sum(\mathrm{X} - \bar{X})^2}{\mathrm{N}}}$$

The square of the standard deviation is called variance. Therefore

Variance $= \sigma^2$

The standard deviation and variance become larger as the square of the data becomes greater. More important, it is readily comparable with other standard deviations and the greater the standard deviation, the greater the variability.

For grouped data, the formula is

$$\sigma = \sqrt{\frac{\sum \mathrm{f}(\mathrm{X} - \bar{X})^2}{\mathrm{N}}}$$

The following formulas for standard deviation are mathematically equivalent to the above formula and are often more convenient to use in calculations.

$$\sigma = \sqrt{\frac{\sum \mathrm{fX}^2}{\mathrm{N}} - \left(\frac{\sum \mathrm{fX}}{\mathrm{N}}\right)^2} = \sqrt{\frac{\sum \mathrm{fX}^2}{\mathrm{N}} - \bar{X}^2}$$

$$= \sqrt{\frac{\sum \mathrm{fd}^2}{\mathrm{N}} - \left(\frac{\sum \mathrm{fd}}{\mathrm{N}}\right)^2} \times \mathrm{i} \text{ Where } \mathrm{d} = \frac{\mathrm{X} - \mathrm{A}}{\mathrm{i}}$$

**Remarks:** If the data represent a sample of size N from a population, then it can be proved that the sum of the squared deviations are divided by (N-1) instead of by N. However, for large sample sizes, there is very little difference in the use of (N-1) or N in computing the standard deviation.

To understand the formula for grouped data, consider the following data which relate to the profits of 100 companies.

| **Profit (Rs. lakhs)** | **No. of companies** | **Profit (Rs. lakhs)** | **No. of companies** |
|---|---|---|---|
| 8-10 | 8 | 14-16 | 30 |
| 10-12 | 12 | 16-18 | 20 |
| 12-14 | 20 | 18-20 | 10 |

To compute standard deviation we construct the following table:

| Profits (Rs. lakhs) | m.p. X | f | d = (X-15)/2 | fd | $fd^2$ |
|---|---|---|---|---|---|
| 8-10 | 9 | 8 | -3 | -24 | 72 |
| 10-12 | 11 | 12 | -2 | -24 | 48 |
| 12-14 | 13 | 20 | -1 | -20 | 20 |
| 14-16 | 15 | 30 | 0 | 0 | 0 |
| 16-18 | 17 | 20 | +1 | +20 | 20 |
| 18-20 | 19 | 10 | +2 | +20 | 40 |
| | | N = 100 | | $\sum fd = -28$ | $\sum fd^2 = 200$ |

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i = \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 2$$

$$= \sqrt{2 - 0.0784} \times 2 = \sqrt{1.9216} \times 2$$

$$= 1.3862 \times 2 = 2.7724 \simeq 2.77$$

The standard deviation is most commonly used to measure variability, while all other measures have rather special uses. In addition, it is the only measure possessing the necessary mathematical properties (like combined standard deviation) to make it useful for advanced statistical work.

**Activity E**

The following data show the daily sales at a petrol station. Calculate the mean and standard deviation.

| Number of litres sold | No. of days | Number of litres sold | No. of days |
|---|---|---|---|
| 700-1000 | 12 | 1900-2200 | 18 |
| 1000-1300 | 18 | 2200-2500 | 5 |
| 1300-1600 | 20 | 2500-2800 | 2 |
| 1600-1900 | | | |

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 4.9 COEFFICENT OF VARIATION

A frequently used relative measure of variation is the coefficient of variation, denoted by C.V. This measure is simply the ratio of the standard deviation to mean expressed as the percentage.

Coefficient of variation $= \text{C.V.} = \frac{\sigma}{\bar{x}} \times 100$ when the coefficient of variation is less in the data, it is said to be less variable or more consistent.

ignou THE PEOPLE'S UNIVERSITY

Consider the following data which relate to the mean daily sales and standard deviation for four regions.

| Region | Mean daily sales (Rs. thousand) | Standard deviation (Rs. thousand) |
|---|---|---|
| 1 | 86 | 10.45 |
| 2 | 45 | 5.86 |
| 3 | 72 | 72 |
| 4 | 61 | 11.32 |

To determine which region is most consistent in terms of daily sales, we shall compute the coefficients of variation. You may notice that the mean daily sales are not equal for each region.

$$\mathrm{C.V.}_1 = \frac{10.45}{86} \times 100 = 12.15; \mathrm{C.V.}_2 = \frac{5.86}{45} \times 100 = 13.02$$

$$\mathrm{C.V.}_3 = \frac{9.54}{72} \times 100 = 13.25; \mathrm{C.V.}_4 = \frac{11.32}{61} \times 100 = 18.56$$

As the coefficient of variation is minimum for Region 1, therefore the most consistent region is Region 1.

**Activity F**

A factory produces two types of electric lamps, A and B. In an experiment relating to their life, the following results were obtained.

| Length of life (in hours) | Type A No. of lamps | Type B No. of lamps |
|---|---|---|
| 500-700 | 5 | 4 |
| 700-900 | 11 | 30 |
| 900-1100 | 26 | 12 |
| 1100-1300 | 10 | 8 |
| 1300-1500 | 8 | 6 |

Compare the variability of the life of the two types of electric lamps using the coefficient of variation.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 4.10 SKEWNESS

The measures of central tendency and variation do not reveal all the characteristics of a given set of data. For example, two distributions may

ignou THE PEOPLE'S UNIVERSITY

have the same mean and standard deviation but may differ widely in the shape of their distribution. Either the distribution of data is symmetrical or it is not. If the distribution of data is not symmetrical, it is called asymmetrical or skewed. Thus skewness refers to the lack of symmetry in distribution.

A simple method of detecting the direction of skewness is to consider the tails of the distribution (Figure I). The rules are:

Data are symmetrical when there are no extreme values in a particular direction so that low and high values balance each other. In this case, mean = median = mode. (see Fig I(a) ).

If the longer tail is towards the lower value or left hand side, the skewness is negative. Negative skewness arises when the mean is decreased by some extremely low values, thus making mean < median < mode. (see Fig I(b) ).

If the longer tail of the distribution is towards the higher values or right hand side, the skewness is positive. Positive skewness occurs when mean is increased by some unusually high values, thereby making mean > median > mode. (see Fig I(c) )

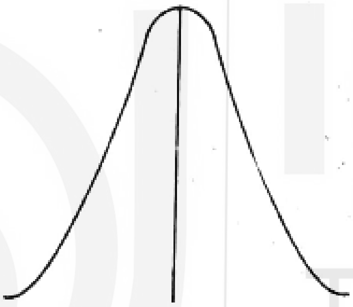


**Fig. 1 (a): Symmetrical**

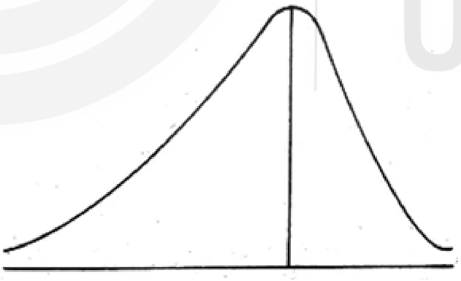


**Fig. 1 (b): Negatively skewed Distribution**

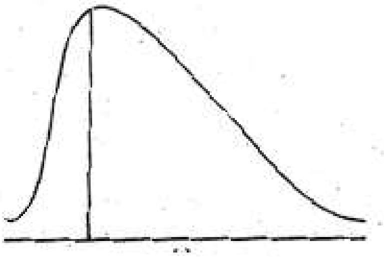


**Fig. 1 (c): Positively skewed distribution**

## 4.11 RELATIVE SKEWNESS

In order to make comparisons between the skewness in two or more distributions, the coefficient of skewness (given by Karl Pearson) can be defined as:

$$\text{SK.} = \frac{\text{Mean - Mode}}{\text{S. D.}}$$

If the mode cannot he determined, then using the approximate relationship, Mode = 3 Median - 2 Mean, the above formula reduces to

$$\text{SK.} = \frac{3\ (\text{Mean - Median})}{\text{S.D.}}$$

if the value of this coefficient is zero, the distribution is symmetrical; if the value of the coefficient is positive, it is positively skewed distribution, or if the value of the coefficient is negative, it is negatively skewed distribution. In practice, the value of this coefficient usually lies between ± 1.

When we are given open-end distributions where extreme values are present in the data or positional measures such as median and quartiles, the following formula for coefficient of skewness (given by Bowley) is more appropriate.

$$\text{SK.} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Again if the value of this coefficient is zero, it is a symmetrical distribution. For positive value, it is positively skewed distribution and for negative value, it is negatively skewed distribution.

To explain the concept of coefficient of skewness, let us consider the following data.

| **Profits (Rs. thousand)** | **No. of companies** | **Profits (Rs. thousand)** | **No. of companies** |
|---|---|---|---|
| 10-12 | 7 | 18-20 | 25 |
| 12-14 | 15 | 20-22 | 10 |
| 14-16 | 18 | 22-24 | 5 |
| 16-18 | 20 | | |

Since the given distribution is not open-ended and also the mode can be determined, it is appropriate to apply Karl Pearson formula as given below:

$$\text{SK.} = \frac{\text{Mean - Mode}}{\text{S. D.}}$$

| Profits (Rs. thousand) | m.p. X | f | d= (X- 17)/2 | fd | $fd^2$ |
|---|---|---|---|---|---|
| 10-12 | 11 | 7 | -3 | -21 | 63 |
| 12-14 | 13 | 15 | -2 | -30 | 60 |
| 14-16 | 15 | 18 | -1 | -18 | 18 |

| | | | | | |
|---|---|---|---|---|---|
| 16-18 | 17 | 20 | 0 | 0 | 0 |
| 18-20 | 19 | 25 | + 1 | 25 | 25 |
| 20-22 | 21 | 10 | +2 | 20 | 40 |
| 22-24 | 23 | 5 | +3 | 15 | 45 |
| | | N = 100 | | $\sum fd = -9$ | $\sum fd^2 = 251$ |

$$\bar{X} = A + \frac{\sum fd}{N} \times i = 17 - \frac{9}{100} \times 2 = 17 - 0.18 = 16.82$$

$$\text{Mode } = \text{L} + \frac{\text{d}_1}{\text{d}_1 + \text{d}_2} \times \text{i} = 18 + \frac{5}{5 + 15} \times 2 = 18 + 0.5 = 18.5$$

$$\sigma = \sqrt{\frac{\sum \text{fd}^2}{\text{N}} - \left(\frac{\sum \text{fd}}{\text{N}}\right)} \times \text{i} = \sqrt{\frac{251}{100} - \left(\frac{-9}{100}\right)} \times 2$$

$$\sqrt{2.51 - 0.0081} \times 2 = \sqrt{2.509} \times 2. = 1.5817 \times 2 = 3.1634$$

$$\text{Sk} = \frac{16.82 - 18.5}{3.1634} = -0.5310$$

This value of coefficient of skewness indicates that the distribution is negatively skewed and hence there is a greater concentration towards the higher profits.

The application of Bowley's method would be clear by considering the following data:

| **Sales (Rs. lakhs)** | **No. of companies** | **c.f.** |
|---|---|---|
| Below 50 | 8 | 8 |
| 50-60 | 12 | 20 |
| 60-70 | 20 | 40 |
| 70-80 | 25 | 65 |
| 80 & above | 15 | 80 |

$Q_1$ = size of $\frac{N}{4}$th observation $= \frac{80}{4}$ = 20th observation which lies in the class 50-60

$$\text{Q}_1 = \text{L} + \frac{\text{N}/4 - \text{pcf}}{\text{f}} \times \text{i} = 50 + \frac{20 - 8}{12} \times 10 = 60$$

$Q_2$ = Median = size of $\frac{N}{2}$th observation $= \frac{80}{2}$ = 40th observation which lies in the class 60-70

$$\text{Q}_2 = \text{Med.} = \text{L} + \frac{\text{N}/2 - \text{pcf}}{\text{f}} \times \text{i} = 60 + \frac{40 - 20}{20} \times 10 = 70$$

$Q_3$ = Size of $\frac{3N}{4}$th observation $= \frac{3 \times 80}{4}$ = 60 th observation which lies in the class 70-80

$$\text{Q}_3 = \text{L} + \frac{3\text{N}/4 - \text{pcf}}{\text{f}} \times \text{i} = 70 + \frac{60 - 40}{25} \times 10 = 78$$

ignou
THE PEOPLE'S
UNIVERSITY

$$\text{Coefficient of } SK = \frac{Q_3 + Q_1 - 2\text{ Median}}{Q_3 - Q_1}$$

$$= \frac{78 + 60 - 2 \times 70}{78 - 60} = -0.11$$

This value of coefficient of skewness indicates that the distribution is slightly skewed to the left and therefore there is a greater concentration of the sales at the higher values than the lower values of the distribution.

## 4.12 SUMMARY

In this unit, we have shown how the concepts of measures of variation and skewness are important. Measures of variation considered were the range, average deviation, quartile deviation and standard deviation. The concept of coefficient of variation was used to compare relative variations of different data. The skewness was used in relation to lack of symmetry.

## 4.13 KEY WORDS

**Average Deviation** is the arithmetic mean of the absolute deviations from the mean or the median.

**Coefficient of Variation** is a ratio of standard deviation to mean expressed as percentage.

**Interquartile Range** considers the spread in the middle 50% ($Q_3 - Q_1$ ) of the data.

**Quartile Deviation** is one half the distance between first and third quartiles.

**Range** is the difference between the largest and the smallest value in a set of data.

**Relative Variation** is used to compare two or more distributions by relating the variation of one distribution to the variation of the other.

**Skewness** refers to the lack of symmetry.

**Standard Deviation** is the root mean square deviation of a given set of data.

**Variance** is the square of standard deviation and is defined as the arithmetic mean of the squared deviations from the mean.

## 4.14 SELF- SSESSMENT EXERCISES

1) Discuss the important of measuring variability for managerial decision making.
2) Review the advantages and disadvantages of each of the measures of variation.
3) What is the concept of relative variation? What problem situations call for the use of relative variation in their solution?
4) Distinguish between Karl Pearson's and Bowley's coefficient of skewness. Which one of these would you prefer and why?
5) Compute the range and the quartile deviation for the following data:

| Monthly wage (Rs.) | No. of workers | Monthly wage (Rs.) | No. of workers |
|---|---|---|---|
| 700-800 | 28 | 1000-1100 | 30 |
| 800-900 | 32 | 1100-1200 | 25 |
| 900-1000 | 40 | 1200-1300 | 15 |

6) Compute the average deviation for the following data:

| No. of shares applied for | No. of applicants | No. of shares applied for | No. of applicants |
|---|---|---|---|
| 50-100 | 2500 | 250-300 | 900 |
| 100-150 | 1500 | 300-350 | 750 |
| 150-200 | 1300 | 350-400 | 675 |
| 200-250 | 1100 | 400-450 | 525 |
| | | 450-500 | 450 |

7) Calculate the mean, standard deviation and variance for the following data

| No. of defects per item | Frequency | No. of defects per item | Frequency |
|---|---|---|---|
| 0-5 | 18 | 25-30 | 150 |
| 5-10 | 32 | 30-35 | 100 |
| 10-15 | 50 | 35-40 | 90 |
| 15-20 | 75 | 40-45 | 80 |
| 20-25 | 125 | 45-50 | 50 |

8) Records were kept on three employees who wrapped packages on sweet boxes during the Diwali holidays in a big sweet house. The study yielded the following data

| Employee | Mean number of packages | Standard deviation' |
|---|---|---|
| A | 23 | 1.45 |
| B | 45 | 5.86 |
| C | 32 | 3.54 |

i) Which package wrapper was most productive?

ii) Which employee was the most consistent?

iii) What measure did you choose to answer part (ii) and why?

9) The following data relate to the mileage of two types of tyre:

| **Life (in kms.)** | **Number of Tyres** | |
|---|---|---|
| | **Type A** | **TypeB** |
| 20000-22000 | 230 | 200 |
| 22000-24000 | 270 | 275 |
| 24000-26000 | 450 | 470 |
| 26000-28000 | 375 | 300 |
| 28000 30000 | 125 | 155 |

i) Which of the two types gives a higher average life?

ii) If prices are the same for both the types, which would you prefer and why?

10) The following table gives the distribution of daily travelling allowance to salesmen in a company:

| **Travelling Allowance (in Rs.)** | **No. of salesmen** | **Travelling Allowance(Rs.)** | **No. of salesmen** |
|---|---|---|---|
| 100-120 | 14 | 180-200 | 15 |
| 120-140 | 16 | 200-220 | 7 |
| 140-160 | 20 | 220-240 | 6 |
| 160-180 | 18 | 240-260 | 4 |

Compute Karl Pearson's coefficient of skewness and comment on its value.

11) Calculate Bowley's coefficient of skewness from the following data:

| Monthly wages | No. of workers | Monthly wages | No. of workers |
|---|---|---|---|
| Below 600 | 10 | 800-900 | 20 |
| 600-700 | 25 | 900-1000 | 15 |
| 700-800 | 45 | 1000 & above | 5 |

12) You are given the following information before and after the settlement of workers' strike.

| | **Before sell lenient of strike** | **After settlement of strike** |
|---|---|---|
| No, or workers | 1000 | 950 |
| Average Wage (Rs.) | 1300 | 1350 |
| Standard Deviation (Rs.) | 400 | 425 |
| Median Wage (Rs.) | 1325 | 1300 |

Assuming that the increase in wage is a loss to the management, comment on the gains and losses from the point of view of workers and that of management.

## 4.15 FURTHER READINGS

Bowerman, B.L., O' Connell, R.T., Business Statistics in Practice, McGraw Hill.

Clark, T.C. and E.W. Jordan. Introduction to Business and Economic Statistics, South-Western Publishing Co.:

Enns, P.G., Business Statistics, Richard D. Irwin Inc.: Homewood.

Gupta, S.P. and M.P. Gupta. Business Statistics, Sultan Chand & Sons: New Delhi.

Moskowitz, H. and G.P. Wright. Statistics for Management and Economics, Charles E. Merill Publishing Company.

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# BLOCK 2

# PROBABILITY AND PROBABILITY DISTRIBUTIONS

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 5 BASIC CONCEPTS OF PROBABILITY

### Objectives

After reading this unit, you should be able to:

- appreciate the relevance of probability theory in decision-making
- understand the different approaches to probability
- calculate probabilities in different situations
- revise probability estimate, if added information is available.

### Structure

5.1 Introduction
5.2 Basic Concepts : Experiment, Sample Space, Event
5.3 Different Approaches to Probability Theory
5.4 Calculating Probabilities in Complex Situations
5.5 Revising Probability Estimate
5.6 Summary
5.7 Further Readings

## 5.1 INTRODUCTION

Uncertainty is a part and parcel of human life. Weather, stockmarket prices, product quality are but some of the areas, where, commenting on the future with certainty becomes impossible. Decision-making in such areas is facilitated through formal and precise expressions for the uncertainties involved. Study of rainfall, spelled out in a form amenable for analysis, may render the decision on water management easy. Intuitively, we see that if there is a high chance of a large quantity of rainfall in the coming year, we may decide to use more water of rainfall for power generation and irrigation this year. We may also take some steps regarding flood control. However, in order to know how much water to release for different purposes, we need to quantify the chances of different quantities of rainfall in the coming year.

Similarly, formal and precise expressions of stockmarket prices and product quality uncertainties, may go a long to help analyse, and facilitate decision on portfolio and sales planning respectively. Probability theory provides us with the ways and means to attain the formal and precise expressions for uncertainties involved in different situations. The objective of this unit is to introduce you to the theory of probability. Accordingly, the basic concepts are first presented, followed by the different approaches to probability measurement that have evolved over time. Finally, in the last two sections, certain important results in quantifying uncertainty which have emerged as a sequel to the theoretical developments in the field, are presented.

ignou THE PEOPLE'S UNIVERSITY

**Activity A**

Mention three events in your life, where you faced total certainty.

1) ……………………………………………………………………………

……………………………………………………………………………

2) ……………………………………………………………………………

……………………………………………………………………………

3) ……………………………………………………………………………

……………………………………………………………………………

**Activity B**

Mention two major events in your life, where you faced uncertainty in taking decisions. Elaborate as to how you dealt with the uncertainty in each of the cases.

1) ……………………………………………………………………………

……………………………………………………………………………

2) ……………………………………………………………………………

……………………………………………………………………………

ignou
THE PEOPLE'S
UNIVERSITY

## 5.2 BASIC CONCEPTS: EXPERIMENT, SAMPLE SPACE, EVENT

Probability, in common parlance, connotes the chance of occurrence of an event or happening. In order that we are able to measure it, a more formal definition is required. This is achieved through the study of certain basic concepts in probability theory, like experiment, sample space and event. In this section we explore these concepts.

**Experiment**

The term experiment is used in probability theory in a much broader sense than in physics or chemistry. Any action, whether it is the tossing of a coin, or measurement of a product's dimension to ascertain quality, or the launching of a new product in the market, constitute an experiment in the probability theory terminology.

These experiments have three things in common:

1) There are two or more outcomes of each experiment.
2) It is possible to specify the outcomes in advance.
3) There is uncertainty about the outcomes.

For example, a coin tossing may result in two outcomes, in head or tail, which we know in advance, and we are not sure whether a head or a tail will come up when we toss the coin. Similarly, the product we are measuring may turn out to be undersize or right size or oversize, and we are not certain which way it will be when we measure it. Also, launching a new product involves uncertain outcome of meeting with a success or failure in the market.

**Sample Space**

The set of all possible outcomes of an experiment is defined as the sample space. Each outcome is thus visualised as a sample point in the sample space. Thus, the set (head, tail) defines the sample space of a coin tossing experiment. Similarly, (success, failure) defines the sample space for the launching experiment. You may note here, that given any experiment, the sample space is fully determined by listing down all the possible outcomes of the experiment.

**Event**

An event, in probability theory, constitutes one or more possible outcomes of an experiment. Thus, an event can be defined as a subset of the sample space. Unlike the common usage of the term, where an event refers to a particular happening or incident, here, we use an event to refer to a single outcome or a combination of outcomes. Suppose, as a result of a market study experiment of a product, we find that the demand for the product for the next month is uncertain, and may take values from 100, 101, 102... 150. We can obtain different event like:

The event that demand is exactly 100

The event that demand lies between 101 to 120

The event that demand is 101 or 102

In the first case, out of the 51 sample points that constitute the sample space, only one sample point or outcome defines the event, whereas the .number of outcomes used in the second and third case are 20 and 2 respectively.

With this background on the above concepts, we are now in a position to formalise the definition of probability of an event. In the next section, we will look at the different approaches to probability that have been developed, and present the axioms for the definition of probability.

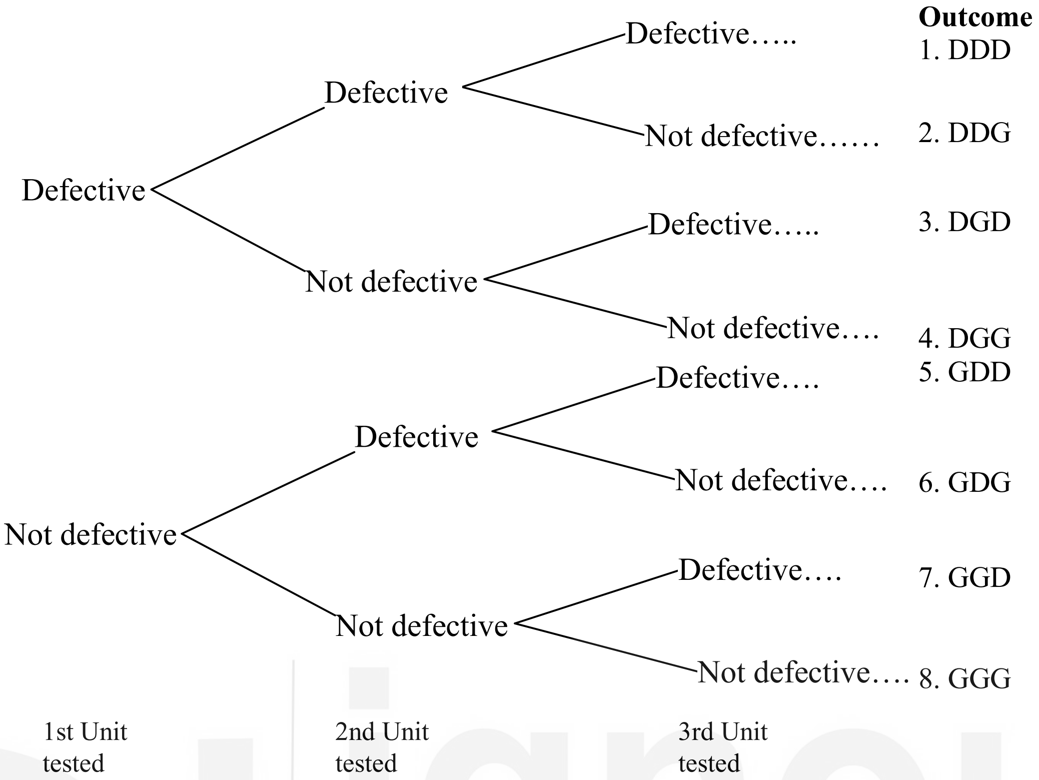
**Example 1**

Consider the experiment of testing three units of a product. We are interested in finding the possible outcomes of this test.

**Solution**

In this experiment, we find that each unit can be either defective or not defective. The test results of the three units may be represented as follows :

ignou
THE PEOPLE'S
UNIVERSITY


Outcome
Defective
Defective
Defective…..
Not defective……
Not defective
Defective…..
Not defective….
Not defective
Defective
Defective….
Not defective….
Not defective
Defective….
Not defective….
1. DDD
2. DDG
3. DGD
4. DGG
5. GDD
6. GDG
7. GGD
8. GGG
1st Unit tested
2nd Unit tested
3rd Unit tested

ignou
THE PEOPLE'S
UNIVERSITY

The above diagram shows all possible outcome (here 8 in number) of the experiment. Corresponding to each of the two outcomes of the testing of one unit, the second unit may be defective or non-defective, leading to 2 x 2 = 4 outcomes. Corresponding to each of these four outcomes, the third unit may again give two results giving us in total 4 x 2 = 8 possible outcomes of the experiment.

If we denote a defective by D and a non-defective by G, then the sample space(s) can be written down as the list of all possible outcomes of the experiment ;

S = (DDD, DDG, DGD, DGG, GDD, GDG, GGD, GGG)

**Example 2**

Suppose we are interested in the following Event A in the above experiment: The number of defective are exactly two. How many sample' points does this event correspond to?

**Solution**

We can see from the sample space that there are three outcomes where D occurs twice, viz, DDG, DGD and GDD, thus the Event A corresponds to 3 sample point.

**Activity C**

Consider an experiment where four coins are tossed once. List down the possible outcomes of the experiment. In how many outcomes do you find the occurrence of two heads?

……………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 5.3 DIFFERENT APPRAOCHCES TO PROBABILITY THEORY

Three different approaches to probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally sought. In this section, we first explore the approaches through examples of distinct types of experiments. The axioms that are common to these approaches are then presented, and the concept of probability is defined using the axioms.

Consider the following situations marked by three distinct types of experiments. The events that we are interested in, within these experiments, are also given.

**Situation 1**

Experiment : Drawing a number from among nine number(s) (say 1 to 9).

Event : On any draw, number 4 occurs.

**Situation 2**

Experiment : Administering a particular drug.

Event : The drug puts a person to sleep in ten minutes.

**Situation 3**

Experiment : Commissioning a solar power plant.

Event : The plant turns out to be a successful venture.

The first situation is characterized by the fact that on any draw, each of the nine members has got equal changes of occurrence, if the experiment is conducted with fairness. Thus, in any draw, any one of the numbers may turn up, and the changes of occurrence of each is equal. This type of situation, marked by the presence of "equally likely" customers, gave rise to the Classical Approach of probability theory. In the Classical Approach, probability of an event is defined as :

$$\frac{\text{the number of outcomes favorable to the event}}{\text{total number of outcomes}}$$

Thus if we denote the event that "a 4 comes out in a draw" as A, and the probability of the event as P (A), then we can see that the total number of outcomes in a draw of numbers 1 to 9 is 9, as any one of the numbers may occur. The number 4 occurs only

Once in these 9 outcomes. Thus $P(A) = \frac{1}{9}$

We no have a look at the second situation. If we try to apply the above definition of probability in the second experiment, we find that we cannot say that the drug will be equally active for all persons. Moreover, we do not know as to how many persons have been tested. This implies that we should have the past date on people who were administered the drug and the number that fell asleep in ten minutes. In the absence of past data we have to undertake an experiment, where we administer the drug on a group of people to check its effect. We are assuming here that experimentation is safe, i.e., the drug does not have any side effects.

The Relative Frequency Approach is used to compute probability in such cases. As per this approach, the probability of occurrence of an event is given by the ratio of the number of times the event occurs to the total numbers of trials. Denoting the event by B, and the probability of the event by P(B), we can write :

$$P(B) = \frac{\text{Number of persons who fell asleep in 10 minutes}}{\text{total number of persons who were given the drug}}$$

It was recognized in this approach that, in order to take such a measure, we should have it tested for a large number of people. In other words, the total number of trials in the experiment should be very large.

The third situation seems apparently similar to the second one. You may be tempted here to apply the Relative Frequency Approach. Denoting by C, the event that the venture is a success, and P(C) the probability of success, you may calculate P(C) as :

$$\frac{\text{Number of successful ventures}}{\text{Total number of such ventures undertaken}}$$

that is, the relative frequency of successes will be a measure of the probability. However, the calculation here presupposes that either, (a) it is possible to do an experiment with such ventures, or (b) that past data on such ventures will be available. In practice, a solar power plant being a relatively new development involving the latest technology, past experiences are not available. Experimentation is also ruled out because of high cost and time involved, unlike the drug testing situation. In such cases, the only way out is the Subjective Approach to probability. In this approach, we try to access the probability from our own experiences. We may bring in any information to access this. In the situation cited, we may perhaps, look into the performance of the commissioning authority in other new and related technologies. You may note here that sine the assessment is a purely subjective one, it will vary from person to person.

**Activity D**

Which approach to probability will you apply in the following situations?

**Situation 1.** Drawing a card from a deck of cards.

……………………………………………………………………………………

……………………………………………………………………………………

**Situation 2.** Selecting a person at random from a group of 50 people.

…………………………………………………………………………………

…………………………………………………………………………………

**Situation 3.** Determining the effect of alcohol on accident.

…………………………………………………………………………………

…………………………………………………………………………………

**Situation 4.** Determining the chances of a third world war.

…………………………………………………………………………………

If you have not faced any problem while attempting the exercise in classifying the above situations, you should try your understanding with the following. Let us assume that we are interested n finding out the chances of getting a head in the toss of a coin. By now, you would have come up with the answer by the classical approach, using the argument, that there are two outcomes, heads and tails, which ar equally likely. Hence, given that a head can occur only once, the probability is $\frac{1}{2}$ Consider the following alternative line of argument, where the probability can be estimated using the Relative Frequency Approach. If we toss the coin for a sufficiently large number of times and note down the number of times the head occurs, the proportion of times that a head occurs will give us the required probability.

Thus, given our definition of the approaches, we find both the arguments to be valid. This bring out, in a way, the commonality between the Relative Frequency and the Classical Approach. The difference, however, is that the probability computed by using the Relative Frequency Approach will be tending to be $\frac{1}{2}$ with a large number of trials; moreover an experiment is necessary in this case. In comparison, in the Classical Approach, we know apriori that the changes are $\frac{1}{2}$, based on our assumption of "equally likely" outcomes. As already noted, the different approaches have evolved to cater to different kinds of situations. To highlight the commonality of the Classical Approach and the Relative Frequency Approach vis-à-vis the third approach, the first two are sometimes referred to as the objectivist's measures in contrast to the subjectivist's measure of the third approach.

All these approaches, however, share the same basic axioms. These axioms are fundamental to probability theory and provide us with unified approach to probability. The axioms are;

A) The probability of an event A, written as P(A), must be a number between zero and one, both values inclusive. Thus $0 \leq P(A) \leq 1$.

B) The probability of occurrence of one of the other of all possible events is equal to one. As S denotes the sample space or the set of all possible events, we write $P(S) = 1$. Thus in tossing a coin once, P(a head or a tail) = 1.

C) If two events are such that occurrence of one implies that the other cannot occur, then the probability that either one or the other will occur is equal to the sum of their individual probabilities. Thus, in a coin tossing situation, the occurrence of a head rules out possibility of occurrence of tail. These events are called mutually exclusive events. In such cases then, if, A and B are the two events respectively, then P( A or B) = P (A) + P (B) i.e. P (Head) + P (Tail).

You may note here that these being axioms, they cannot be proved. It follows from the last two axioms that if two mutually exclusive events form the sample space of the experiment, then P (A or B) = P(A) + P(B) = 1; thus P (Head) +P (Tail) = 1. If two or more events together define the total sample space, the events are said to be collectively exhaustive.

Given the above axioms, we may not define probability as a function which assigns probability value P to each sample point of an experiment abiding by the above axioms. Thus, the axioms themselves define probability. We are not in a position to look into the process of calculating probabilities for more complex situations. In the next section, we attempt this through a study of the probability theorems.

**Example 3**

Suppose, we are interested in finding out the probability that the toss of three coins will result in only heads. We reason out the possible outcomes of the experiment are :

(A) All heads, (B) two heads and one tail, (C) one head and two tails and (D) all tails. Since there are four possible outcomes and only one result with all heads, the required probability is $\frac{1}{4}$.

Detect the error in our argument.

**Solution**

We can see clearly that the total number of possible outcomes of the experiment is 8, the sample space being :

(HHH, HHT, HTH, HTT, TTT, TTH, THT, THH)

Out of the above 8 possible outcomes only one outcome is favorable to the event of all heads, hence the probability is $\frac{1}{8}$, if we were to list down the four different outcomes as reasoned out, we find that the event listed above occur as follows :

Event A All heads : once, HHH

Event B Two heads and one tail : three times, namely, HHT, HTH, THH.

Event C Two tails and one head : three times, namely TTH, THT, HTT.

Event D All tails : Once, TTT

Thus, if all the possible outcomes are classified as above, then it follows that the changes of occurrence of the above event are $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$ respectively, i.e. the event A,B,C,D are not equally likely, as we have taken in our argument. Hence the error.

**Example 4**

A persons who sells newspaper wants to find out the changes that on any day he will be able to sell more than 100 copies of Indian Express. From his diary where he has recorded the daily sales of last years, he find out that out of 365 days, on 73 days he had sold 85 copies, on 146 days he had sold 95 copies, on 60 days, he had sold 105 copies and on 86 days he had sold 110 copies of Indian Express. Can you help him to find out the required probability?

**Solution**

Taking Relative Frequency Approach we find :

| Sale | No. of days (Frequency) | Relative Frequency |
|---|---|---|
| 85 | 73 | 73/365 |
| 95 | 146 | 146/365 |
| 105 | 60 | 60/365 |
| 110 | 86<br>365 | 86/365 |

Thus the number of days when his sales were more than 100 = (60 + 86) days = 146 days. Hence the required probability = $\frac{146}{365}$ = 0.4

**Activity E**

Calculating the probability of drawing an ace from a deck of 52 cards

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

**Activity F**

A proof reader is interested in finding probability that the number of mistakes in a page will be less than 10 from his past experience be finds that out of 3000 pages he has proofread, 100 pages contained no errors, 900 pages contained 5 errors, and 2000 pages contained 12 or more errors Can you help him in finding the probability?

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

ignou THE PEOPLE'S UNIVERSITY

## 5.4 CALCULATING PROBABILITIES IN COMPLEX SITUATIONS

In the last section, we have seen how to compute probabilities in certain situations. The nature of the events were relatively simple, so that direct application of the definition of probability could be used for computation. Quite often, we are interested in the probability of occurrence of more complex events. Consider for example, that you want to find the probability that an ace or a spade will occur in a draw from a deck of 52 cards. You may be interested in the occurrence of this event because you have betted on the same. Similarly, on examining couples with two children, if one child is known as a boy, you may be interested in the probability of the event of both the children being boys. These two situations, we find, are not as simple as those discussed in the earlier section. As a sequel to the theoretical development in the field of probability, certain results are available which help us in computing probabilities in such situations. In this section, we explore these results through examples.

**Example 5**

Suppose that you have taken the examinations on all the three courses given in Module 1 by IGNOU. You have received the following information about the results of your batch. All your batch mates have appeared for all the three courses and,

35% have failed in course 1,
20% have failed in course2,
25% have failed in course3,
10% have failed in both course 1 and 2,
5% have failed in both courses 1 and 3,
8% have failed in both course 2 and 3, and
2% have failed in all the three courses.
You are interested in finding out the probability that any one of your batch mates passing in all the three courses.

**Solution**

A pictorial representation of the problem helps immensely in solving such probabilities. The representation is called a Venn Diagram. In a Venn Diagram the whole sample space of an experiment is represented by a rectangle and different events are visualized as different areas inside the rectangle. The same Venn Diagram area can be used to represent the probability space itself, with probability of occurrence of the rectangle as 1 (being the sample space), and probabilities of events as areas inside the rectangle.

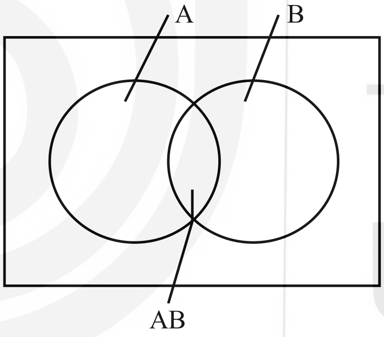
Thus, if two events have an overlap, they will be shown as interesting with one another, while two mutually exclusive events, by definition being non-overlapping, will never interest.

We can now try to represent the given problem through a Venn Diagram. We define the following events.

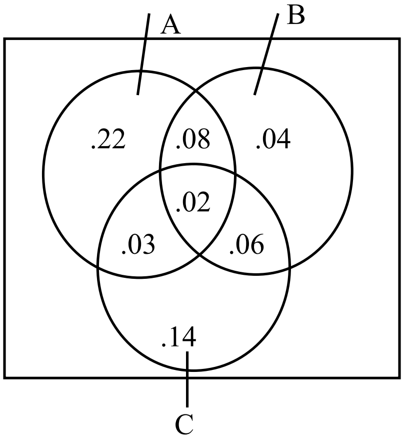
A : Failure in course 1
B : Failure in course 2
C : Failure in Course 3
AB : Failure in courses 1 and 2
AC : Failure in courses 1 and 3
BC : Failure in courses 2 and 3
ABC : Failure in all the courses

The probabilities of the above events are given by the relative frequency approach as P(A) = 0.35; P(B) = 0.2; P(C) = 0.25 and so on.

A rectangle of unit 1 is first drawn. It represents the probability of the sample space of the experiment, namely, results of the three courses. Circle A, with area = 0.35 inside this rectangle that represents P(A). If we are to draw another circle, B, of area 0.2 representing P(B), we find that A and B should interest, as the two events A and B are overlapping. The information tells us that there are some people who have failed in both courses 1 and 2 (event AB). The value of P(AB) = 0.10 gives us the area that is common to both A and B. Therefore, circle B is to be drawn intersecting A, so that the overlap area between the circles is 0.10. We have then the following diagram :



How do we now draw the third circle? We find that C has overlapping areas with both A and B, as there are instances of failure in courses 1 and 3 (AC) and 2 and 3 (BC). There is an instance of failure in all the subjects (ABC) also. Thus, the circle C can be drawn as follows :



Each circle in the process is divided into four areas, the value of each is shown inside the respective areas in the diagram.

The values are derived as follows :

P(ABC) represents the area common to the events A, B and C i.e. failure in all the subjects, and is given by 02. We also know that P(AB) i.e. the area common to A and B is 10. Thus P(AB) represents failure in courses 1 and 2 and as such contains people who have failed in courses 3 also…….

**Similarly**

Prob. of failure in course 1 and 3 but not in 2 = .05 – .02 = .03

Prob. of failure in courses 2 and 3 but not in 1 = .08 – .02 = .06

We know the failure percentage in course 1, given by P(A) = .35 is divided into four segments with three segments having area of .08, .02 and .03. These area basically mean that out of the people who have failed in course 1, some have failed in co0urse 2 also but not in 3, some have failed in course 3 but not in course2, and some have failed in courses 2 and 3. Hence the remaining area of the circle A, will be (.35 – .08 – .02 – .03) = 0.22 representing the probability of failure only in course 1.

Similarly, we find the other two areas as .04 and .14 (as shown in the diagram) representing the probabilities of failures in course 2 only and course 3 only respectively. The total area enclosed by all the circles can be found by adding up areas of all the segments:

$$.22 + .08 + .02 + .03 + .06 + .14 + .04 = 0.59$$

The balance area in the rectangle is then (1-0.59) = 0.4 and gives the probability of the event of “ pass in all the three courses”.

(you may do well to list down all the mutually exclusive and exhaustive events of the experiment of results of courses 1 , 2 and 3).

The required probability that any person will pass in all the subjects is 0.41.

**Example 6**

Consider your locality, where, out of the 5000 people residing, 1200 are above 30 years of age and 3000 are female. Out of the 1200 who are above 30, 200 are female. Suppose, after a person is chosen you are told that the person is a female. What is the probability that she is above 30?

**Solution**

We define the following events :

A : a person chosen is above 30 years.

B: a person chosen is female.

We are interested in the event A, given that B has occurred. If we denote this event by A/B, we want to find P(A/B).

Out of the 1200 persons who are above 30, 200 are females.

Out of 5000 people in the locality, 200 possess the characteristics of being female as well as above 30 years. Using a notation similar to the last example, we4 define the event AB as :

AB : Event that a person is both a female and above 30 years of age we derive from the data given that :

$$P(A)=\frac{1200}{5000},\ \ P(B)=\frac{3000}{5000}\quad \text{and } P(AB)=\frac{200}{5000}$$

To find the probability that, given a female has been chosen, she will be above 30, we see that out of 3000 females in the total population only 200 females are above 30. Thuis the required probability is

$$\frac{200}{3000}\quad \text{i.e. } P(A/B)\quad \frac{200}{3000}.$$

You may note here that

$$P(A/B)=\frac{200}{3000}=\frac{200}{5000}\times\frac{5000}{3000}=\frac{200}{5000}\Big/\frac{3000}{5000}=\frac{P(AB)}{P(B)}$$

The examples given above refer to probability computations under two distinct types of situations :

1) In the first example, we were interested in calculating probabilities for events which are not mutually exclusive.
2) In the second example, we wanted to find the probability of an event given that

Such situations have long since been formalized and probability theory and we have results in this area that may be used directly to calculate the required probability in such situations. We present the results here without proof.

**Result 1.** The probability of occurrence of either A or B, written as P (A or B), is given by P(A) + P (B) – P(AB), where

P(A) = Probability of A occurring
P(B) = Probability of B occurring
P(AB) = Probability of A and B jointly occurring.

P(AB) is referred to as the joint probability and P(A) and P(B) are the unconditional or marginal probabilities of the events.

**Result 2.** The probability of occurrence of either A or B or C, written as P(A or B or C), is given by P(A) + P(B) + P(C) – P(AB) – P(BC) – P(CA) + P (ABC). This result is directly applicable to Example 5, and the notation here are the same as those used in the example.

Thus, P(A or B or C) basically gives the area enclosed by all the circles and represents the probability that a person will fail in either course 1 or course 2 course 3.

Thus, in Example 5, we found P (A or B or C) = 0.59

Now, if we denote the event D = pass in all three subjects, then the sample space of the experiment can be divided into two mutually exclusive and exhaustive events (1) D and (2) A or B or C.

Hence, Probability of the sample space = P(S) = P (D) + P(A or B or C) = 1 or ; P(D) = 1 - .59 = .41

**Result 3.** If the occurrence of event B affects the probability of occurrence of event A, then the probability of occurrence of A, given that B has occurred (known as conditional probability), denoted by P(A/B) is equal to

$$P(A/B) = \frac{P(AB)}{P(B)} \qquad \text{or } P(AB) = P(A/B) \times P(B)$$

This has been demonstrated in Example 6.

A special case of this result is when the events A and B are independent, i.e. the occurrence of B does not affect the probability of occurrence of A. Then P(A/B) = P(A); hence P(AB) = P(A) × P(B)

An example of independent events is the successive tossing of coins. Hence the probability of two heads occurring in two successive tosses

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The above summarizes certain key results in probability theory which help us in calculating probabilities in relatively complex situations. In practice, we often find that after we have estimated the change of occurrence of an event, added information on the event becomes available to us, based on which we revise the odds. In the final section, we present the theory behind such revision of probabilities.

**Activity G**

What is the probability of getting an ace or a spade in a draw from a deck of 52 cards?

………………………………………………………………………………………

………………………………………………………………………………………

**Activity H**

Consider families with only two children, and you are told that in a family one of their children is a boy, what is the probability that the other child is also a boy?

………………………………………………………………………………………

………………………………………………………………………………………

## 5.5 REVISING PROBABILITY ESTIMATE

As we have already noted in the introduction, the basic objective behind calculating probabilities to help us in making decisions. Quite often, whether it is in our personal life or our work life, decision making is an ongoing process. Consider for example, a seller of winter garments. You are interested in the demand of the product. To help you in deciding on the amount you should stock for the winter, you have computed the probability of selling different quantities and have noted that the change of selling a large quantity is very high. Accordingly, you have taken the decision to stock a large quantity of the product. Suppose, when finally the winter comes and the season ends, discover that you are left with a large quantity of stock. Assuming that you are in this business, you fell that the earlier probability calculation should be updated given the new experience to help you decide on the stock for the next winter. Similarly situations exist where we are interested in an event on an ongoing basis. Every time some new information is available, we do revise our odds mentally. This revision of probability with added information is formalized in probability theory in terms of a theorem known as Bayes' Theorem. In the final section of this unit, we present the Bayes' Theorem for revising probability estimates.

**Bayes' Theorem**

It states that if A and B are two mutually exclusive and collectively exhaustive events and C is another event defined in the context of the same experiment, then given the values of P(A), P(B)(, P(C/A) and P(C/B), the conditional probability

$$P(A/C) = \frac{P(C/A) \times P(A)}{P(C/A) \times P(A) + P(C/B)\,P(A)}$$

$$\text{and } P(B/C) = \frac{P(C/B).P(B)}{P(C/B)\ P(B) + P(C/A)\,P(A)}$$

We discuss now an application of the above theorem.

Consider the case of a manufacturer who is using a particular machine for producing a product. From earlier data, he has estimated that the chances of the machine being set correctly or incorrectly are 0.6 and 0.4 respectively. Thus, we have two mutually exclusive and collectively exhaustive events:

A: The set up is correct

B: The set up is incorrect

With P(A) = 0.6 and P(B) = 0.4 (check P(A) + P(B) = 1, follows from definition mutually exclusive and collectively exhaustive events).

From the past data, the manufacturer has estimated that when a machine is set correctly, it produces 10% defectives, otherwise, it produces 60% defectives.

Suppose, before going for the batch production, on unit has been produced, inspected and found to be defective. The manufacturer wants to find out in what way this new information affects the probabilities of the events A and B. In other words, given added information, what is the probabilities of the events A and B. In other words, given added information, what is the probability that the set up is correct?

AC : A defective occurs with set up correct

BC : A defective occurs with set up incorrect

P (C) = P(AC) + P(BC) …….. (1)

We also have the data on conditional probabilities

P (Defective/set up correct) = P(C/A) = 0.1

P (Defective/set up incorrect) = P(C/B) = 0.6

From result 3 of the last section, we know :

P(AC) = P(C/A) P(A) and

P(BC) = P (C/B) P (B)

Therefore (1) can be written as :

P (C)= P(C/A) P(A) + P(C/B) P (B)

We are interested in finding the probability that the set up was correct given that a defective piece has come out i.e. P(A/C)

Again from result 3, $P(A/C) = \frac{P(AC)}{P(C)} = \frac{P(C/A)\ P(A)}{P(C/A)\ P(A) + P(C/B)\ P(B)}$

Hence the proof of the Bayes. Theorem

$$\therefore P(A/C) = \frac{0.10 \times 0.6}{0.1 \times 0.6 + 0.6 \times 0.4} = \frac{.06}{.06 + .24} = 0.2$$

Thus, with the new information we have revised the probability estimate of a correct set up. As the new piece was defective, this has reduced the probability of correct set up.

We can consider a further variation in the above situation. Suppose instead of basing his calculation on a single piece, the manufacturer produced two pieces before going in for batch production. If he found both the pieces defective, what is the probability that the current set up is correct?

Now the event C is defined as both pieces being defective, P(C/A) denotes the probability that given the set up was correct, both pieces have turned out to be defective.

The change of one defective with a correct set up is 10%

$\therefore$ P (C/A) = .1 × .1 = .01

Similarly P(C/B) = .6 × .6 = .36

Applying the theorem directly,

$$P(A/C) = \frac{P(C/A) \times P(A)}{P(C/A)P(A) + P(C/B)P(B)} = \frac{.01 \times .6}{.01 \times .6 + .36 \times .4} = .04$$

The above formula demonstrate the method of using added information to revise the probability estimates.

With this section, we are in a position to calculate probabilities under different situations. However, we have not demonstrated explicitly as to how exactly this probability information can be used for decision-making. Before we do this, certain other concepts and standard representations in probability theory need to be discussed. We take up these issue in the subsequent units.

**Activity I**

Instead of two mutually exclusive and collectively exhaustive events, as considered in the above text, suppose that there are 3 events A1, A2 and A3 which are mutually exclusive and collectively exhaustive. Defining the event C in the same way as in the text, show that

$$P(AI/C) = \frac{P(C/A1)\,P(A1)}{P(C/A1)\,P(A1) + P(C/A2)\,P(A2) + P(C/A3)\,P(A3)}$$

Derive the expression for P(A2/C) and P(A3/C)

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity J**

Find a real life analogy of the situation given in Activity I and explain the use fullness of such probability calculation in decision making.

(**Hind :** A1, A2, A3 may respectively be the events of having an undersize, right size and oversize product. Undersize pieces of the product is much less desirable as compared to oversize ones, as reworking is possible in the latter, while the former will be scrap.)

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 5.6 SUMMARY

Probability in common parlance means the change of occurrence of an event. The need to develop a formal and precise expression for uncertainty in decision-making, has led to different approaches to probability measurement. These approaches, namely, the Classical, the Relative Frequency and the Subjectivists' Approach, arose mainly to cater to different types of situations where we face uncertainties. The approaches, however, share the same basic axioms. In this unit, we have used these axioms to define probability formally and the definition has been used to calculate probabilities of different types of events. A|s the events, of interest to us, become more complex, the computation of probability through definition turns out to be tedious. Certain results in probability theory which are helpful in this context have been presented. The need to revise the odds in the light of new information is felt in many situations. In the final section of this unit, we have shown the method to revise the probability estimate as added information on the outcome of the experiment becomes available.

## 5.7 FURTHER READINGS

Chance, W., Statistical methods for Decision Making, R. Irwin Inc: Homewood

Feller, W., An Introduction to Probability Theory and its Applications, John Wiley & Sons Inc.: New York

Levin, R., Statistics for Management, Prentice-Hall Inc.: Englewood-Cliffs.

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 6 DISCRETE PROBABILITY DISTRIBUTIONS

**Objectives**

After reading this unit, you should be able to :

- understand the concepts of random variable and probability distribution
- appreciate the usefulness of probability distribution in decision-making
- identify situations where discrete probability distributions can be applied
- find or assess discrete probability distributions for different uncertain situations
- appreciate the application of summary measures of a discrete probability distribution.

**Structure**

6.1 Introduction
6.2 Basic Concepts : Random Variable and Probability Distribution
6.3 Discrete Probability Distributions
6.4 Summary Measures and their Applications
6.5 Some Important Discrete Probability Distributions
6.6 Summary
6.7 Further Readings

## 6.1 INTRODUCTION

In our study of Probability Theory, we have so far been interested in specific outcomes of an experiment and the chances of occurrence of these outcomes. In the last unit, we have explored different ways of computing the probability of an outcome. For example, we know how to calculate the probability of getting all heads in a toss of three coins. We recognise that this information on probability is helpful in our decisions. In this case, a mere 0.125 chance of all heads may dissuade you from betting on the event of "all heads". It is easy to see that it would have been further helpful, if all the possible outcomes of the experiment together with their chances of occurrence were made available. Thus, given your interest in betting on head's, you find that a toss of three coins may result in zero, one, two or three heads with the respective probabilities of $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}$, and $\frac{1}{8}$. The wealth of information, presented in this way, helps you in drawing many different inferences. Looking at this information, you may be more ready to bet on the event that either one or two heads occur in a toss of three coins. This representation of all possible outcomes and their probabilities is known as a probability distribution. Thus, we refer to this as the probability distribution of "number of heads" in the experiment of tossing of three coins. While we see that our previous knowledge on computation of probabilities helps us in arriving at such representations, we recognise that the calculations may be quite tedious. This is apparent, if you try to calculate the probabilities of different number of heads in a tossing of twelve coins. Developments in Probability Theory help us in specifying the probability distribution in such cases with relative ease. The theory also gives certain standard probability distributions and provides the conditions under which they can be applied. We will study the probability distributions and their applications in this and the subsequent unit. The objective of this unit is to look into a type of probability distribution, viz., a discrete probability distribution. Accordingly, after the initial presentation on the basic concepts and definitions, we will discuss as to how discrete probability distributions can be used in decision-making.

**Activity A**

Suppose you are interested in betting on 'tails' in a tossing of four coins. Write down the result of the experiment in terms of the "number of tails" (zero to four) that may occur, with their respective probabilities of occurrence. Elaborate as to how this may help you in betting.

………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

## 6.2 BASIC CONCEPTS : RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

Before we attempt a formal definition of probability distribution, the concept of ‘random variable' which is central to the theme, needs to be elaborated.

In the example given in the Introduction, we have seen that the outcomes of the experiment of a toss of three coins were expressed in terms of the "number of heads" Denoting this "number of heads" by the letter H, we find that in the example, H can assume values of 0, 1, 2 and 3 and corresponding to each value, a probability is associated. This uncertain real variable H, which assumes different numerical values depending on the outcomes of an experiment, and to each of whose values a probability assignment can be made, is known as a random variable. The resulting representation of all the values with their probabilities is termed as the probability distribution of H. It is customary to present the distribution as follows :

**Probability Distribution of Number of Heads (H)**

| H | P(H) |
|---|---|
| 0 | 0.125 |
| 1 | 0.375 |
| 2 | 0.375 |
| 3 | 0.125 |

In this case, as we find that H takes only discrete values, the variable H is called a discrete random variable and the resulting distribution is a discrete probability distribution.

In the above situation, we have seen that the random variable takes a limited number of values. There are certain situations where the variable of interest may take infinitely many values. Consider for example that you are interested in ascertaining the probability distribution of the weight of the one kilogram tea pack, that is produced by your company. You have reasons to believe that the packing process is such that the machine produces a certain percentage of the packs slightly below one kilogram and some above one kilogram. It is easy to see that there is essentially to chance that the pack will weigh exactly 1.000000 kg., and there are infinite number of values that the random variable "weight" can take. In such cases, it makes sense to talk of the probability that the weight will be between two values, rather than the probability of the weight will be between two values, rather than the probability of the weight taking any specific value. These types of random variables which can take an infinitely large number of values are called continuous random variables, and the resulting distribution is called a continuous probability distribution. Sometimes, for the sake of convenience, a discrete situation with a large number of outcomes is approximated by a continuous distribution: Thus, if we find that the demand of a product is a random variable taking values of 1, 2, 3... to 1000, it may be worthwhile to treat it as a continuous variable. Obviously, the representation of the probability distribution for a continuous random variable is quite different from the discrete case that we have seen. We will be discussing this in a later unit when we take up continuous probability distributions.

Coming back to our example on the tossing of three coins, you must have noted the presence of another random variable in the experiment, namely, the number of tails (say T). T has got the same distribution as H. In fact, in the same experiment, it is possible to have some more random variables, with a slight extension of the experiment. Supposing a friend comes and tells you that he will toss 3 coins, and will pay you Rs. 100 for each head and Rs. 200 for each tail that turns up. However, he will allow you this privilege only if you pay him Rs. 500 to start with.

You may like to know whether it is worthwhile to pay him Rs. 500. In this situation, over and above the random variables H and T, we find that the money that you may get is also a random variable. Thus,

if H =number of heads in any outcome, then 3 - H = number of tails in any outcome (as the total number of heads and tails that can occur in a toss of three coins is 3)

The money you get in any outcome = 100H + 200 (3 - H)

$$= 600 - 100H = x \text{ (say)}$$

We find that x which is a function of the random variable H, is also a random variable.

We can see that the different values x will take in any outcome are

$$(600 - 100x0) = 600$$

$$(600 - 1010 \text{ x } 1) = 500$$

(600-100 x 2 =400

(600-100 x 3) =300

Hence the distribution of x is :

| X | p(X) |
|---|---|
| 600 | $\frac{1}{8}$ |
| 500 | $\frac{3}{8}$ |
| 400 | $\frac{3}{8}$ |
| 300 | $\frac{1}{8}$ |

The above gives you the probability of your getting different sums of money. This may help you in deciding whether you should utilise this opportunity by paying Rs. 500.

From the discussion on this section, it should be clear by now that a probability distribution is defined only in the context of a random variable or a function of a random variable. Thus in any situation, it is important to identify the relevant random variable and then find the probability distribution to facilitate decision-making.

In the next section we will look at the properties of discrete probability distributions and discuss the methods for finding and assessing such distributions.

**Activity B**

Suppose three units of a product are tested. The result of the test is given in terms of pass or fail. If the probability that a unit will pass inspection is 0.8, find the probability distribution of the number of units that pass inspection.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 6.3 DISCRETE PROBABILITY DISTRIBUTIONS

In the previous section we have seen that a representation of all possible values of a discrete random variable together with their probabilities of occurrence is called a discrete probability distribution. The objective of this section is to look into the properties of such distributions, and discuss the methods for assessing them.

In discrete situations, the function that gives the probability of every possible outcome is referred to in Probability Theory as the "probability mass function" (p.m.f.).The outcomes, as you must have noted, are mutually exclusive and collectively exhaustive Thus, a representation of the p.m.f. of the number of heads H, in a toss of three coins can be :

$$\begin{array}{r|l} f(H) = 0.125'' & \text{if } H = 0 \text{ heads} \\ = 0.375 & \text{if } H = 1 \text{ heads} \\ = 0.375 & \text{if } H = 2 \text{ heads} \\ = 0.125 & \text{if } H = 3 \text{ heads} \end{array}$$

Thus, we see that p.m.f. is the name given to a discrete probability distribution, and if, for any situation, we can specify the p.m.f. of the relevant random variable, the whole probability distribution is then specified. The properties of any p. m. f. , say f(x) where x the random variable, can be derived from the fact that f(x) basically refers to probability values. Any probability measure is by definition non-negative $\sum f(x)=1$ Moreover, it follows from probability theory, that f(x) = 1, the sum being taken over all the possible outcomes.

Sometimes, we are interested in finding the probability of a group of outcomes. In such cases, an addition of the relevant values gives us the result. Thus, in the example given earlier, we find that the probability of 2 or 3 heads = f(2) + f(3) = .5. Further, we may be interested in the probability that the random variable will take values less than or equal to a particular quantity. The result in such situations is achieved by specifying what is known as cumulative distribution function (c.d.f.). The c.d.f. denoted by F(H) is formed by adding the probabilities up to a given quantity, and it gives the probability that the random variable H will take a value less than or equal to that quantity. The F(H) in the example discussed earlier can be written as :

$$F(H) = \begin{cases} 0.125 & \text{for } H = 0 \\ 0.05 & \text{for } H = 1 \text{ or less} \\ 0.875 & \text{for } H = 2 \text{ or less} \\ 1.0 & \text{for } H = 3 \text{ or less} \end{cases}$$

we can see from the above c.d.f. that the probability of getting 2 or less heads is 0.875.

Assessment of the p.m.f. of a random variable follows directly from the different approaches to probability that we have discussed in the earlier unit. The different methods by which p.m.f. of a random variable can be specified are :

1) using standard functions in probability theory
2) using past data on the random variable
3) using subjective assessment.

We now discuss each of the methods and the situations where these can be applied.

**Using Standard Functions**

Sometimes the knowledge of the underlying process in an experiment helps us to specify the probability mass function. Probability theory has come out

ignou
THE PEOPLE'S
UNIVERSITY

with standard functions and the conditions under which these standard functions can be applied to any experiment. Consider again the p.m.f. for the random variable H in the tossing of three coins. An alternative way of specifying f(H) would be as follows :

$$\mathrm{f(H)} = \frac{3!}{\mathrm{H}!(3-\mathrm{H})!}\left(\frac{1}{2}\right)^{\mathrm{H}}\left(\frac{1}{2}\right)^{3-\mathrm{H}}, \text{ for H} = 0,1,2,3.$$

Where H! is read as H factorial and is given by:

$$\mathrm{H}! = 1 \times 2 \times 3 \times \dots\dots\dots\dots \times \mathrm{H}. \text{ and } 0! = 1$$

Thus, for H = 0, we have $f(0) = \frac{3!}{0!3!}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{3} = \frac{1}{8}$

Similarly, you can verify that the values you get for f(1), f(2), f(3) by substituting 1, 2 and 3 in the above function, are the same as obtained those obtained earlier.

This form of f(H) is made possible, as the coin tossing experiment satisfies the conditions specific to a **Bernoulli Process**. Bernoulli Process is defined in probability theory as a process marked by dichotomous outcomes with probability of an event remaining constant from trial to trial. In coin tossing, we find that the outcome of any toss is either a head or a tail, so that the dichotomy is preserved. Also in each of three coin tosses, the probability of head (or tail) remains constant, namely $\frac{1}{2}$. The probability distribution pertaining to such a process is standardised in probability theory, so that we can directly write down the p.m.f. corresponding to any experiment that satisfies the Bernoulli Process. Such standard discrete distributions will be discussed in detail in a later section.

**Using Past Data**

Past data on the variable of interest is used to assess the p.m.f., only if we have reasons to believe that conditions similar to the past will prevail. The frequency of occurrence of each of the values of the variable are noted down and the relative frequency of each of the values is taken as a probability measure. The basis lies in the Relative Frequency Approach discussed in the last unit. You may like to compare the resulting p.m.f. with the corresponding frequency distribution. Thus, under the assumption that buyer behaviour has not changed much, we take the past sales data of a product to find the probability distribution of future sales. While frequency distribution is simply a representation of what has happened in the past, p.m.f. represents what we can expect in the future. If you refer now to Example 4 of the last unit, you can see that the probability distribution of the random variable "daily sales of Indian Express has been estimated from past data. If we denote the random variable by x, we can write down the p.m.f. as :

$$f(x) = \begin{cases} 73/365 \text{ for } x = 85 \\ 146/365 \text{ for } x = 95 \\ 60/365 \text{ for } x = 105 \\ 86/365 \text{ for } x = 110 \end{cases}$$

This method of assessing the p.m.f. stems from the Subjectivists Approach to probability. This method is applied if there is no past data, and the situation of interest does not resemble any known processes in Probability theory. Suppose a record manufacturing company is contemplating the introduction of a new **ghazal** singer. Before introducing him, they want to find out the likely sales of record of the new person in the first year of the release of the record. The random variable here is the "sales in first year". Let us denote it by S. We may here use our subjective assessment to find the p.m.f. of S. One way to assess this may be as follows. The company knows that currently one lakh people buy their records and it believes that out of this one lakh people, 20% i.e. 20,000 customers have the attitude to try anything new, so that the other 80,000 will never buy an unknown singer's record in the first year of release. They have also assessed that at least 10% of their customers are always ready for new ghazals. Building up on such assessments, the final p.m.f. of S may be:

$$\mathrm{f(s)} = \begin{cases} .6 \text{ for } \mathrm{S} = 10{,}000 \\ .2 \text{ for } \mathrm{S} = 15{,}000 \\ .2\text{for}\mathrm{S} = 20{,}000 \end{cases}$$

In other words, they expect that sales in the first year will be 10,000 with a 60% chance, and 20% chance each that 15,000 or 20,000 people will buy it.

We have seen the different ways to assess a discrete probability distribution. These distributions help us in our decisions by presenting the total scenario in an uncertain situation. The p.m.f. of sales as discussed above, may help the company in deciding how many records should be produced in the first year. While producing 10,000 records is definitely a safe thing to do, we realise that a 40% chance of not being able to meet demand is also there. Similarly production of 20,000 records takes care of meeting all demands that may arise, but then there is a chance that some records may not be sold. Systematic analysis of such decisions can be done with the p.m.f. and the relevant cost data, and will be taken up in Unit 8. Analysis is made easier, if together with the p.m.f. data, certain key figures of the p.m.f. are presented. Thus, it may be easier for us to see things, if the expected sales figure is given to us in the above case. These key figures pertaining to a p.m.f. are called summary measures. In the next section we discuss some summary measures that are helpful in analyzing situations.

**Activity C**

Cheek whether the following p.m.f. applies for the random variable in activity B

$$f(X) = \frac{3!}{X!\,(3-X)!}(.8)^{x}(.2)^{3-x}$$

where X = the number of units that pass inspection

(Hint : find f(0), f(1), f(2) and f(3) by substituting X = 0, 1, 2, and 3 in the above function. Check whether these values are the same as what you obtained earlier.)

…………………………………………………………………………………

…………………………………………………………………………………

## 6.4 SUMMARY MEASURES AND THEIR APPLICATIONS

As the name implies, a summary measure of a probability distribution basically summarizes the distribution through a single quantity. Just as we have seen in the case of a frequency distribution, here too we have the measure of location and dispersion that help us to have a quick picture of the behaviour of the random variable concerned. The objective of this section is to look into some of the summary measures and discuss the possible application of these measures.

**Measures of Location**

The most widely used location measure is the Expected Value. It is similar to the concept of mean of a frequency distribution and is calculated as the weighted average of the values of the random variable, taking the respective probabilities of occurrence as the weight. Thus, in the tossing of three coins, the Expected Value of Number of Heads, written as E(H) can be found as follows :

$$\mathrm{E(H)} = \sum \mathrm{H} \times \mathrm{f(H)} = 0 \times .125 + 1 \times .375 + 3\mathrm{x}.125 = 1.5$$

Similarly, considering the extension of the experiment as discussed earlier, we can calculate the money you can expect if you take up your friend's proposal, as :

$$\mathrm{E(X)} = 600\mathrm{x}.125 + 500 \times .375 + 400 \times .375 + 300 \times .125 = \mathrm{Rs}.450$$

Recalling that you have to pay Rs. 500 to get the privilege of entering this game, you may decide not to go in for it as the expected pay off is less than the sum you have to pay. It may be noted in this context that the pay off X at any outcome is a function of the random variable H. As already noted, X itself is a random variable. Instead of calculating the E(X) as above, it is possible to calculate the E(X) as follows :

$$\mathrm{E(X)} = \mathrm{E}(600 - 100\mathrm{H}) = 600 - 100\mathrm{E(H)} = 600 - 100 \times 1.5 = 450$$

It can be seen that for any linear function g(H) of H, the following holds : E[g(H)] = g[E(H)]. That this is not true, for functions other than linear can be verified by taking, for example, g(H) = $H_2$

$$\mathrm{E(H^2)} = \sum \mathrm{H^2 f(H)} = 0 \times .125 + 1 \times .375 + 4 \times .375 + 9\mathrm{x}.125 = 3$$

$$\mathrm{However[E(H)]^2} = (1.5)^2 = 2.25$$

$$\mathrm{Thus\ [E(H)]^2 \# E(H^2)}.$$

Expected value of a random variable gives us a measure of location and is an indicator of the long-run average value that we can expect. In the computation of the expected value, the most likely outcome is given the highest weight age. Sometimes, it is useful to characterize the probability distribution by the most likely value, which is defined as the mode. The modal value is the value corresponding to which, the probability of occurrence is maximum. Another met Sure of location that is of interest is known as 'fractal'. A value Hz is defined as the k fractal of the distribution of H, if

$F(H) < k$ for all $H < H_z$

and $F(H) \geq k$ for all $H \leq H_z$

Recalling the c.d.f. of H, we have developed earlier

| H | F(H) | |
|---|---|---|
| 0 | 0.125 | |
| 1 | 0.500 | |
| | | 0.60 |
| 2 | 0.875 | |
| 3 | 1.000 | |

Suppose we want to find the .60th fractal of the distribution, i.e., we want to find a value of $H = H_k$ such that $F(H) < .60$ for $H \leq H_k$ and $F(H) \geq .60$ for all $H \geq Hk$. We identify that .60 lies between .50 and .875 F(H) values. This is shown by an arrow in the above distribution. The value of H just above it is one that will be the .60th fractile H = 2 is the required answer. We can verify that for H < 2 i.e. for H = 0 and 1, F (0) = .125 and F(1) = .5, both of which are less than 0.6. Similarly for all H ≥ 2, F(2) = .875 and F(3) = 1, both of which are greater than .60. Hence it satisfies the conditions.

You may note that the .50th fractile here is 1, i.e. if any required fractile coincides with any F(H) value in the distribution then the value with which it matches, is the required value. You may verify whether this satisfies the stated conditions. The .5th fractile is called the median of the distribution and is of interest at times.

**Measures of Dispersion**

Standard Deviation (SD), range and absolute deviation are the measures of dispersion of a distribution. Of these, SD being the most widely used, we will discuss it here. You may recall that the same term has been used in the context of a frequency distribution also. However, in a discrete probability distribution, we are dealing with a random variable, and the distribution represents various values of the random variable that we expect will occur in the future. In such, cases, the variance is defined as the expected value of the square of the difference between the random variable and its expected value. Then SD is given by the square root of the variance. Thus, for the random variable H in the coin tossing example, we can write :

$$\text{Variance} = \mathrm{E}[\mathrm{H} - \mathrm{E(H)}]^2 = \mathrm{E}[\mathrm{H} - 1.5]^2$$

$$= (0 - 1.5)^2 f(0) + (1 - 1.5)^2 f(1) + (2 - 1.5)^2 f(2) + (3 - 1.5)^2 f(3)$$

$$= 2.25 \times \frac{1}{8} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{3}{8} + 2.25 \times \frac{1}{8} = \frac{3}{4}$$

and $\text{S.D.} = \sqrt{\text{variance}} = \frac{\sqrt{3}}{2}$

The knowledge on expected value and standard deviation of a distribution of a random variable is useful in our decisions. Suppose you have got an offer to take up any one of the two projects A and B. Both A and B have got uncertain outcomes, so that the payoff for A and B are random variables. If expected payoff for project A is equal to that of project B, and S. D. of payoff in the case of A is less than that of B, then you may decide to choose project A. Here S.D. summarises the variability in monetary payoffs that we can expect from the projects.

We now take up an example to illustrate the use of expected value in decision¬making. More complex situations will be taken up later when we study Decision Theory.

**Example 1**

Consider a newspaper seller who gets newspapers from the local office of the Newspaper every morning and sells them from his shop. He buys each copy for 60 p. and sell it for Rs. 1.10p. However, he has to tell the office in advance as to how many copies he will buy. The office takes back the copies he is not able to sell and pays him only 30 p. for each copy. His problem is essentially to find out how many copies he should order every day. He has estimated the p.m.f. of the daily demand from past data

$$f(D) = \begin{cases} 0.1 \text{ for } D = 30 \\ 0.2 \text{ for } D = 31 \\ 0.2 \text{ for } D = 32 \\ 0.3 \text{ for } D = 33 \\ 0.1 \text{ for } D = 34 \\ 0.1 \text{ for } D = 35 \end{cases}$$

**Solution**

To analyse such situations, first we formalise the problem in terms of alternative courses of actions open to the newspaper man. As he expects that the daily demand will not be less than 30 or more than 35, we understand that there is no point in his ordering less than 30 or more than 35 copies. Thus, he has got six options :

Alternative 1. Order 30 copies

Alternative 2. Order 31 copies

Alternative 3. Order 32 copies

Alternative 4. Order 33 copies

Alternative 5. Order 34 copies

Alternative 6. Order 35 copies

Corresponding to each alternative action, there are six possible values that the demand can take and each of these values lead to a monetary payoff with different chances of occurrence. We can calculate the expected monetary payoff fat each alternative and choose the alternative that promise us the highest expected payoff.

For calculating monetary payoff corresponding to any outcome and any action, we note:

1) If he orders X copies and demand (D) turns out to be more than or equal to X, then he will be able to sell only X copies, so that the payoff will he (1-10 - 0.60) x X = 0.50 X
2) If he orders X copies and D turns out to be less than X, then he will be able to sell D copies for which he will profit 0.5 D and he will be losing (.60 - .30) = 30 p. for each copy he ordered more, i.e. loss = .30 (X-D).

$$\text{His payoff} = .5D - .3X + .3D$$
$$= .8D - .3X$$

With the above background, we are now in a position to calculate the payoff P corresponding to each outcome of an alternative. As these payoff values correspond to the demand values only, the chances of occurrence of the payoffs are given by the chances of occurrence of the respective demand figures. Thus, for each alternative, the p.m.f. of P and the corresponding Expected value of P can be calculated. A sample calculation for Alternative 4 (order 33 copies) is shown below.

**Alternative 4.**

Order 33 copies (X = 33)

| **Outcome** | **Demand(D)** | **If D≥X then P=.5 X If D<X then P=.8D – .3X** | **P** | **f(P)** |
|---|---|---|---|---|
| 1 | 30 | P=.8x30-.3x33 | 14.1 | .1 |
| 2 | 31 | P=.8x31-.3x33 | 14.9 | .2 |
| 3 | 32 | P = .8 x32 - .3x33 | 15.7 | .2 |
| 4 | 33 | P=.5x33 | 16.5 | .3 |
| 5 | 34. | P=.5x33 | 16.5 | .1 |
| 6 | 35 | P=.5 x33 | 16.5 | .1 |

$$\begin{aligned} E(P) &= 14.1x.1 + 14.9x.2 + 15.7x.2 + 16.5x.3 + 16.5x.1 + 16.5 \times .1 \\ &= 1.41 + 2.98 + 3.14 + 4.95 + 1.65 + 1.65 = 15.78 \end{aligned}$$

Similarly, we can calculate the Expected payoff for other alternatives also. The newspaper man should go for the alternative that gives him the highest expected payoff A convenient representation of the alternatives and the outcomes is given below. Corresponding to alternative 4, we have filled up the values. You may now fill up the other cells.

| Probabilities of Demand | | .1 | .2 | .2 | .3 | .1 | | .1 | Expected Payoff E(P) |
|---|---|---|---|---|---|---|---|---|---|
| Order (Alternative) | Demand (Outcomes) | 30 | 31 | 32 | 33 | 34 | 35 | | |
| 1. | 30 | | | | | | | | |
| 2. | 31 | | | | | | | | |
| 3. | 32 | | | | | | | | |
| 4. | 33 | 14.1 | 14.9 | 15.7 | 16.5 | 16.5 | 16.5 | 15.78 | |
| 5. | 34 | | | | | | | | |
| 6. | 35 | | | | | | | | |

ignou
THE PEOPLE'S
UNIVERSITY

On solving E(P), we find that the maximum expected payoff is obtained for Alternative 4. Hence we can say that the newspaper man should order for 33 copies.

**Activity D**

In the above problem, instead of calculating the payoffs, we could have calculated the expected opportunity loss for each alternative.

We recognise that for each alternative and an outcome, three situations can arise:

1) Number ordered (X) = Number demanded (D) : In this case there is no loss to the newspaper man as he has stocked the right number of copies.
2) Number ordered (X) < Number demanded (D) : In this case, he has understocked. and for each copy that he has not ordered for and could have sold, he loses the profit = 0.50 p. Thus, opportunity loss = .50 (D-X).
3) Number ordered (X) > Number demanded (D) : In this case he has ordered for more than he can sell, so he loses (.60-.30) = .30 p. for each extra copy that he has ordered therefore opportunity loss = 0.30 (X-D).

Using the above, calculate the opportunity loss corresponding to each outcome of each alternative. Find the Expected opportunity loss for each alternative and state how you will decide on the basis of these expected values.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 6.5 SOME IMPORTANT DISCRETE PROBABILITY DISTRIBUTIONS

While examining the different ways of assessing p.m.f., we have noted that proper identification of experiments with certain known processes in Probability theory helps us in writing down the probability distribution function. Two such processes are the Bernoulli and the Poisson. The standard discrete probability distribution that are consequent to these processes are the Binomial and the Poisson distribution. The objective of this final section is to look into the conditions that characterise these processes, and examine the standard distributions associated with the processes. This will enable us to identify situations for which these distributions apply.

**Bernoulli Process**

Any uncertain situation or experiment that is marked by the following three properties is known as a Bernoulli Process.

1) There are only two mutually exclusive and collectively exhaustive outcomes in the experiment.
2) In repeated observations of the experiment, the probabilities of occurrence of these events remain constant.
3) The observations are independent of one another.

Typical examples of Bernoulli process are coin-tossing and success-failure situations. In repeated tossing of coins, for each toss, there are two mutually exclusive and collectively exhaustive events, namely, head and tail. We also know that the probability of a head or a tail remains constant (= $\frac{1}{2}$) from toss to toss, and result of one toss does not effect the result of any other toss.

Similar dichotomy is preserved in testing of different pieces of a product. Each piece when tested may be defective (a failure) or non-defective (a success). We know that the production process is such that the probability of a non-defective in any trial is P and that of a defective = q = (1 - p)

Once the process has stabilised, it is reasonable to assume that the success and failure of each piece is independent of the other and also the probability of a success (p) or a failure (q) remains constant from trial to trial. Thus, it satisfies the conditions of a Bernoulli process.

The random variables that may be of interest in the above situations are :

1) The number of successor failure in a specified number of trials, given the knowledge on the probability of a success in trial. This implies that if the experiment is observed n times then given that the probability of a success is same for any observation, we are interested in finding out the distribution of number of successes that may occur in n observations.
2) The number of trials needed to have a specified number of successes, given the knowledge on the probability of success in any trial. We are interested in finding out the probability distribution of the number of trials required to get a specified number of successes.

The Binomial distribution and the Pascal distribution provide us with the required p.m.fs. in the above two cases. We discuss these two distributions with examples.

**Binomial Distribution**

Let us take the example of a machining process which produces on an average 80% good pieces. We are interested in finding out the p.m.f. of the number of good pieces in 5 units produced from this process. From our definition, this situation is a Bernoulli process, with the probability of success = p = 0.8

Probability of failure or defective pieces = q =1 - p = 0.2.

The number of trials = 5.

Let n be the random variables of interest, i.e. the number of good pieces. As N = 5, obviously r can take values of 0, 1, 2, 3, 4, 5, i.e. as 5 pieces are produced, at the best all 5 can be good pieces. We can now try to calculate the probabilities for different values of r using the results given in the last unit :

r = 0 means all 5 are failure. As the probability of failure is q in every trial, and the trials ,We independent, probability of 5 failures = q x q x q x q x q = $q^5$. The total number of outcomes in the experiment are $2^5$ and we find that only in one outcome all 5 are failures.

Therefore $f(0) = q^5$

r = 1 implies that there is one success and four failures. The probability of this is $pq^4$ However, out of the $2^5$ possible outcomes, one success and four failures can occur in the following ways :

1st unit is a success and the rest are failure i.e. SFFFF

2nd unit is a success and the rest are failure i.e. FSFFF

3rd unit is a success and the rest are failure i.e. FFSFF

4th unit is a success and the rest are failure i.e. FFFSF

5th unit is a success and the rest are failure i.e. FFFFS

where S denotes a success and F a failure. Thus, 1 success and 4 failures can occur in 5 different ways, for each of which the probability is $pq^4$

Hence $f(1) = 5\ pq^4$. Similarly for r = 2, the probability of 2 successes and 3 failures is $p^2\ q^3$. To find the number of outcomes in which 2S and 3F will occur we can use the following. Basically, we want to know the different ways in which $^2$S and $^3$F can be put in a sequence. This is represented by $^5C_2$ read as "five C two" and given by

$^5C_2 = \frac{5!}{3!2!} = 10$

Hence $f(2) = 10p^2q^3$

The required p.m.f. of r is then

$$\mathbf{f(r)} = \begin{cases} q^5 \text{ for } r = 0 \\ 5p^4 \text{ for } r = 1 \\ 10p^2q^3 \text{ for } r = 2 \\ 10p^3q^2 \text{ for } r = 3 \\ 5p^4 \cdot q \text{ for } r = 4 \\ p^5 \text{ for } r = 5 \end{cases}$$

Each of the terms for r = 0 …… 5 correspond to the binomial expansion of $(q + p)^5 = q^5 + 5pq^4 + 10p^2q^3 + 10p^3q^2 + 5p^4q + p^5$, hence the above distribution is known as Binomial distribution.

In general, as Binomial distribution gives the probability of r successes in n trials as

$f(r) = {}^nC_r\ p^r q^{n-r}$

where, ${}^nC_r = \frac{n!}{r!(n-r)!}$, r = 0, 1, 2, 3, …….n

p = probability of success in any trial

q = probability of failure in any trial = 1-p.

often f(r) is written as f(r/n, p ), as n and p are given.

We can verify that the above has got the properties of a p.m.f. We can write down directly the p.m.f. as above for any situation that satisfies the earlier stated conditions.

Given the standard expression, it is possible to calculate the expected value (referred to as the mean) and the variance of a Binomial distribution :

Expected value (Mean) = $\Sigma r f(r) = \Sigma r \cdot \frac{n!}{r!n-r!} p^r q^{n-r}$

As $n! = n \times n - 1 \times n - 2 \ldots\ldots \times\cdot 1 = n(n-1)!$

$r! = r(r-1)!$

and $\Sigma \frac{(n-1)!}{(r-1)!(n-r)!} p^{r-1} q^{n-r} = 1$ being sum of probabilities of all outcomes of number of successes in n - 1 trials.

Mean = $\sum n \frac{(n-1)!}{(r-1)!(n-r)!} p \cdot p^{r-1} q^{n-r}$

$$= \text{np}\ \Sigma \frac{(n-1)!}{(r-1)!\,(n-r)'} p^{r-1} q^{n-r} = np$$

The variance of the distribution can be shown to be “npq”.

As, n, p, q, are given constants for a particular distribution, the mean and variance are also constant. These (n,p) are called parameters of a distribution and are often used to specify a distribution.

**Pascal Distribution**

Suppose we are interested in finding the p.m.f. of the number of trials (n) required to get 5 successes, given the probability p, of success in any trial.

We see that 5 successes can be obtained only in 5 or more trials. Thus, we want to find f(n) for n = 5, 6 …………….. etc.

If n trials are required to get 5 successes then the last trial has to result in a success, while in the rest of the n-1 trials, 4 successes have been obtained. This implies that : f(n) = (probability of 4 successes in n-1 trials) X p.

$= {}^{n-1}C_4 p^4 q^{n-5} \cdot p$

It is customary to write f(n) as f(n/r, p), as r and p are given here. The above satisfies the properties of a p.m.f. The mean and the variance of the distribution are $\frac{r}{p}$ and $\frac{rq}{p^2}$ respectively.

Of the many standard discrete distributions, we have so far discussed the Binomial and the Pascal. We now present the Poisson distribution which is applicable to events occurring randomly over time and space. This p.m.f. has been used widely to represent distributions of several random variables like demand for spare parts, number of telephone calls per hour, number of defects per metre in a bale of cloth, etc. In order to apply this p.m.f. in any situation, the conditions of a Poisson process need to be satisfied. We

ignou
THE PEOPLE'S
UNIVERSITY

discussed these conditions and the Poisson distribution in the following paragraphs.

**Poisson Process and Poisson Distribution**

Conditions specific to the Poisson process are easily seen by establishing them in the context of the Bernoulli process. Let us consider a Bernoulli process with n trials and the probability of success in any trial $= \frac{\mathrm{m}}{\mathrm{n}}$, where $\mathrm{m} \geq 0$. Then we do now that the probability of r successes in n trials is given by:

$$f(r) = {}^{\mathrm{n}}\mathrm{C}_{\mathrm{r}}\left(\frac{m}{n}\right)^{r}\left(1-\frac{1n}{n}\right)^{n-r}$$

We note that $^{\mathrm{n}}\mathrm{C}_{\mathrm{r}} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)(n-2)(n-r+1)(n-r)!}{r!(n-r)!}$

$$\mathrm{f(r)} = \frac{\mathrm{m^r}}{\mathrm{r!}}\frac{\mathrm{n}}{\mathrm{n}} \times \frac{\mathrm{n}-1}{\mathrm{n}} \times \frac{\mathrm{n}-2}{\mathrm{n}} \times \ldots\ldots \frac{(\mathrm{n}-\mathrm{r}+1)}{\mathrm{n}} \times \left(1-\frac{\mathrm{m}}{\mathrm{n}}\right) \times \left(1-\frac{\mathrm{m}}{\mathrm{n}}\right)^{\mathrm{n}}$$

Now, if $\mathrm{n} \to \infty$, then the terms, $\frac{\mathrm{n}-1}{\mathrm{n}}, \frac{\mathrm{n}-2}{\mathrm{n}}, \cdots\cdots \frac{\mathrm{n}-\mathrm{r}+1}{\mathrm{n}}$

and $\left(1-\frac{m}{n}\right)^{-r}$ will all be tending to 1.

Also from a theorem in Calculus, it is known that $\left(1-\frac{m}{n}\right)^{n}$ tends to $\mathrm{e}^{-\mathrm{m}}$, if $\mathrm{n} \to \infty$. Thus, we have $\mathrm{f(r)} = \frac{\mathrm{e}^{-\mathrm{m}} \times \mathrm{m}^{\mathrm{r}}}{\mathrm{r!}}$

The above function is a Piosson p.m.f. Thus, a Poisson process corresponds to a Bernoulli process with a very large number of trials (n) and with a very low probability of success (m/n) in any trial. We will now demonstrate a real life analogy of such a process.

Consider the occurrence of any uncertain event over time or space in such a way that the average occurrence of the event over unit time or space is m. We may take the number of accidents occurring over a time period with m denoting the average number of accidents per month; or we may be interested in the number of defects occurring in a strip if cloth manufactured by a mill, with m denoting the average number of defects per metre. For each of such situations, we see the possibility of dividing the time or space interval into n very small segments such that within a small segment the conditions of the Bernoulli process hold. Thus, one month can be divided into (say) 30 x 24 x 60 intervals of one minute each, so that the probability of occurrence of an accident in any

minute $= \frac{\mathrm{m}}{30 \times 24 \times 60}$, and reduces to a very small quantity, so that there is almost no chance of having two accidents occurring in one minute, The independence property of the Bernoulli trial also holds true here, as a one minute interval basically corresponds to a trial. Similar possibilities also exist in the cloth example.

The above enables us to calculate the probability that r accidents will occur, from the Poisson formula derived earlier. As we have made n very large, and

ignou
THE PEOPLE'S
UNIVERSITY

p very small, and have also verified that the Bernoulli conditions are satisfied, we can write $f(r) = \frac{e^{-m}m^r}{r!}$

as the required p.m.f. in such a cases.

The p.m.f. is alternatively written as f(r/m).

Suppose we want to find the distribution of the number of accidents r, given that there are, on an average, 3 accidents per month. We can find this by putting r = 0, 1, 2, 3, 4, ………… in f(r/3)

$$f(0/3) = \frac{e^{-3} \times 3^0}{0!} = e^{-3} = .0498.$$

The mean and variance of a Poisson distribution are equal and are given by m. This property is sometimes used to check whether the Poisson applies for the event under study.

**Activity E**

A plane has got 4 engines. The probability of an engine failing is 1/3 and each engine may fail independently of the other engine. Find the probability that all the engines will fail. Write down the p.m.f. of 'Failed Engines'

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

**Activity F**

If 1% of the bolts produced by a certain machine are defective, find the probability that in a random sample of 300 bolts, all bolts are good.

[Hint : This is a case of a Binomial distribution with n = 300 and p = .01. We have to find f (0/300, .01). As n is large (300) and p is small (.01), Poisson can be used to calculate the required probability. Poisson with m = np = 300 × .01 = 3 will lead to the answer, i.e., find f(0/3).]

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

**Activity G**

From past experience a Proof reader has found that after he proofreads, there remain 2 errors on an average in a page. What is the probability of finding a page without any error?

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

## 6.6 SUMMARY

We have introduced the concepts of random variable and probability distribution in this unit. In any uncertain situation, we are often interested in the behaviour of certain quantities that take different values in different outcomes of the experiments. These quantities are called random variables and a representation that specifies the possible values a random variable can take, together with the associated probabilities, is called a probability distribution, The distribution of a discrete variable is called a discrete probability distribution and the function that specifies a discrete distribution is called a probability mass function (p.m.f.). We have looked into situations that gives rise to discrete probability distributions, and discussed how these distributions are helpful in decision-making. The concept and application of expected value and other summary measures for such distributions have been presented. Different methods for assessing such distributions have also been discussed. In the final section certain standard discrete probability distributions and their applications have been discussed.

## 6.7 FURTHER READINGS

Gangolli, R.A. and D. Ylvisaker, Discrete Probability, Harcourt, Brace & World, Inc.: New York.

Levin, R.I., Statistics for management, Prentice-Hall, Inc. : Englewood-Cliffs.

Parzen,E., Modern Probability Theory and its Applications, Wiley: New York.

Nick T. Thomopoulos. Probability distributions, Springer.

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 7 CONTINUOUS PROBABILITY DISTRIBUTIONS

## Objectives

After reading this unit, you should be able to:

- identify situations where continuous probability distributions can be applied
- appreciate the usefulness of continuous probability distributions in decision making.
- analyse situations involving the Exponential and the Normal distributions.

## Structure

7.1 Introduction
7.2 Basic Concepts
7.3 Some Important Continuous Probability Distributions
7.4 Applications of Continuous Distributions
7.5 Summary
7.6 Further Readings

## 7.1 INTRODUCTION

In the last unit, we have examined situations involving discrete random variables and the resulting probability distributions. Let us now consider a situation, where the variable of interest may take any value within a given range. Suppose that we are planning for release of water for hydropower generation and irrigation. Depending on how much water we have in the reservoir viz. whether it is above or below the "normal" level, we decide on the amount and time of release. The variable indicating the difference between the actual reservoir level and the normal level, can take positive or negative values, integer or otherwise. Moreover, this value is contingent upon the inflow to the reservoir, which in turn is uncertain. This type of random variable which can take an infinite number of values is called a continuous random variable, and the probability distribution of such a variable is called a continuous probability distribution. The concepts and assumptions inherent in the treatment of such distributions are quite different from those used in the context of a discrete distribution. The objective of this unit is to study the properties and usefulness of continuous probability distributions. Accordingly, after a presentation of the basic concepts, we discuss some important continuous probability distributions, which are applicable to many real-life processes. In the final section, we discuss some possible applications of these distributions in decision-making.

ignou THE PEOPLE'S UNIVERSITY

**Activity A**

Give two examples of a continuous random variables. Note down the difficulties you face in writing down the probability distributions of these variables by proceeding in the manner explained in the last unit.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 7.2 BASIC CONCEPTS

We have seen that a probability distribution is basically a convenient representation of the different values a random variable may take, together with their respective probabilities of occurrence. The random variables considered in the last unit were discrete, in the sense that they could be listed in a sequence, finite or infinite. Consider the following random variables that we have taken up in Unit 10 :

1) Demand for Newspaper (D)
2) Number of Trials (N) required to get r successes, given that the probability of a success in any trial is P.

In the first case, D could take only finite number of integer values, 30, 31, 35;

whereas in the second case, N could take an infinite number of integer values r, r + 1, r + 2 ……….. $\infty$. In contrast to these situations, let us now examine the example cited in the introduction of this unit. Let us denote the variable, "Difference between normal and actual water level", by X. We find that X can take any one of innumerable decimal values within a given range, with each of these values having a very small chance of occurrence. This marks the difference between the continuous variable X and the discrete variables D and N. Thus, in case of a continuous variable, the chance of occurrence of the variable taking a particular value is so small that a totally different representation of the probability function is called for. This representation is achieved through a function known as "probability density function" (p.d.f.). Just as a p.m.f. represents the probability distribution of a discrete random variable, a p.d.f. represents the distribution of a continuous random variable. Instead of specifying the probability that the variable X will take a particular value, we now specify the probability that the variable X will lie within an interval. Before discussing the properties of a p.d.f., let us study the following example.

**Example 1**

Consider the experiment of picking a value at random from all available values between the integers 0 and 1. We are interested in finding out the p.d.f. of this value X. (Alternatively, you may consider the line segment 0-1, with the origin at 0. Then, a point picked up at random will have a distance X from the origin. X is continuous random variable, and we are interested in the distribution of X.)

**Solution**

Let us first try to find the probability that X takes any particular value, say, .32.

The Probability (X = .32), written as P(X = .32) can he found by noting that the 1st digit of X has to be 3, the 2nd digit of X has to be 2 and the rest of the digits have to be zero. The event of the 1st digit having a particular value is independent of the 2nd digit having a particular value, or any other digit having a particular value.

Now, the probability that first digit of X is 3 = $\frac{1}{10}$ (As there are 10 possible numbers 0 to 9).

Similarly the probabilities of the other digits taking values of 2, 0, 0 ...etc. are $\frac{1}{10}$ each.

$\mathrm{P(X = .32)} = \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \approx 0$ ………………(1)

Thus, we find that for a continuous random variable the probability of occurrence of any particular value is very small. Therefore we have to look for some other meaningful representation.

We now try to find the probability of X taking less than a particular value, say .32. Then P(X < .32) is found by noting the following events :

A) The first digit has to be less than 3, or

B) The first digit is 3 but the second digit is less than 2.

$\mathrm{P(X < .32)} = \frac{3}{10} + \frac{1}{10} \times \frac{2}{10} = .32$……………… (2)

Combining (1) & (2) we have :

$\mathrm{P(X \leq .32)} = .32$

Similarly, we can find the probability that X will lie between any two values a and b, i.e., P(a ≤ x ≤ b); this is the type of representation that is meaningful in the context of a continuous random variable.

**Properties of a p.d.f.**

The properties of p.d.f. follow directly from the axioms of probability discussed in Unit 9. By definition, any probability function has to be non-negative and the sum of the probabilities of all possible values that the random variable can take, has to be 1. The summation for continuous variables is made possible through 'integration'.

If f(X) denotes the pdf of a continuous random variable X, then

1) f(x) = 0, and

2) $\int_R f(X)dX = 1$, where " $\int_R$ " denotes the integration over the entire range (R) of values of X.

The probability that X will lie between two values a and b, will be given by : $\int_a^b f(X)dx$.

The cumulative density function (c.d.f.) is found by integrating the p.d.f. from the lowest value in the range upto an arbitrary level X. Denoting the c.d.f. by F(X), and the lowest value the variable can take by a, we have :

$$F(X) = \int_a^x f(X)dx$$

Once the p.d.f. of a continuous random variable is known, the corresponding c.d.f. can be found. You may once again note, that as the variable may take any value in a specified interval on a real line, the probabilities are expressed for intervals rather than for individual values, and are obtained by integrating the p.d.f. over the relevant interval.

**Example 2**

Suppose that you have been told that the following p.d.f. describes the probability of different weights of a "1kg tea pack" of your company :

$$f(x) = \begin{cases} 100(x-1), 1 \le x \le 1.1 \\ 0 \text{ otherwise otherwise} \end{cases}$$

Verify whether the above is a valid p.d.f.

**Solution**

As, $f(x) = 100(x-1)$ for $1 \le x \le 1.1$

= 0 otherwise.

The relevant limits for integration are 1 and 1.1; for all other values below 1 and above 1.1, the probability being zero.

In order that f(x) is a valid p.d.f., two conditions need to be satisfied. We test them one by one.

1 Check f(x) $\ge 0$

i.e. to show that $100(\mathrm{x}-1) \ge 0$ for $1 \le \mathrm{x} \le 1.1$

It is easy to see that this is true, for all other values of x, f(x) is given to be 0. So this condition is satisfied.

2 Check f(x) dx = 1

i.e. To show that $\int_1^{1.1} 100(x-1)dx = 1$

Left Hand side = $100\frac{x^2}{2}\Big|_1^{1.1} - 100x|_1^{1.1}$ (By integration)

$$= \frac{100}{2}[1.1^2 - 1^2] - 100[1.1-1]$$

$$= 50 \times 2.1 \times .1 - 100 \times .1 = 10.5 - 10 = .5$$

As this is not equal to 1, this is not a valid p.d.f.

**Example 3**

The p.d.f. of the different weights of a "1kg tea pack" of your company is given by :

$f(x) = 200(x-1)$ for $1 \leq x \leq 1.1$

= 0, otherwise.

You may note that the packing process is such that even if you set the machine to a value, you will only get packs around that value. The p.d.f. shows that there are chances of only exceeding the 1 kg value and there is no chance of packing less than 1kg. This is normally achieved by setting the machine to a relatively high value to overcome the government regulation on packing standard weights.)

Verify that the given p.d.f. is a valid one. Find the probability that the weight of any pack will lie between 1.05 and 1.10.

**Solution**

Proceeding in the same way as in the earlier example, we can show that

$$\int_{1}^{1.1} 200(x-1)dx = 1$$

Now, we find the probability that x will lie between 1.05 and 1.10 :

$$\mathbf{P}(1.05 \leqslant \mathbf{x} \leqslant 1.10) = \int_{1.05}^{1.10} 200(x-1)dx$$

$$= 100x^2|_{1.05}^{1.10} - 200x|_{1.05}^{1.10}$$

$$= 100(1.1^2 - 1.05^2) - 200(1.1 - 1.05)$$

$$= 100 \times 2.15 \times .05 - 200 \times .05 = 15 \times .05 = .75$$

Alternatively, we could have found the above as follows :

$$P(1.05 \leqslant x \leqslant 1.10) = P(1 \leqslant x \leqslant 1.1) - P(1 \leqslant x \leqslant 1.05)$$

$$= 1 - [100 \times 2.05 \times .05 - 208 - .05]$$

$$= 1 - .25 = .75$$

**Example 4**

find the cdf for the pdf given in Example 3.

**Solution**

$$F(X) = \int_{1}^{x} 200(x-1)dx$$

$$= (100x^2 - 100) - 200(x-1)$$

$$= 100(x^2 - 2x + 1)$$

(Here , 1 is the lowest possible value that x can take).

In this section we have elaborated on the concept of a continuous random variable and have finally shown how to arrive at a representation of the probability function of such a variable. We have used "integration" for our purpose. Those of you who are not familiar with the concept of integration, may note that this is similar to the summation sign (I) used in the context of a discrete variable. Also, if f(x) vs x is plotted on a graph, we will have a curve. The integration between two values a and b of x then signifies the area under the curve, and as we have already seen, this is nothing but the probability that x will lie between a and b. This idea will be useful again when we discuss some important theoretical probability distributions for continuous variables in the next section.

**Activity B**

Suppose that you are told that the time to service a car at your friend's petrol station is uncertain with the p.d.f. given as :

$f(x) = 3x - 2x^2 + 1$ for $0 \leqslant x \leqslant 2$

$= 0$ otherwise.

Examine whether this is a valid p.d.f.

(You may need to brush up Integration from any elementary Calculus book.)

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity C**

The life in Hours of an electric bulb is known to be uncertain. A particular manufacturer of bulbs has estimated the p.d.f. of "life" (the total time for which the bulb will burn before getting fused) as :

f(x) = 0, for x< 0

$= \frac{1}{100} e^{-(x/100),}$ for $x \geq 0$

Check whether the above is a valid p.d.f.

If it is a valid p.d.f., find the probability that a bulb will have a life of more than 100 hours.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 7.3 SOME IMPORTANT CONTINUOUS PROBABILITY DISTRIBUTIONS

The knowledge of the probability density function (p.d.f.) of a continuous random variable is helpful in many ways. The p.d.f. allows us to calculate the probability that a variable will lie within a certain range. The usefulness of such calculations are illustrated with the help of the following two situations.

**Situation 1**

Mr. X manufactures tea and sells it in packets of 1kg. He knows that the packing process is imperfect, so that there is always a chance that any packet that is finally sold will have a tea content exceeding 1kg or less than 1 kg. In the current process, it is possible to set the packing machine, so that the packet weighs within a certain range. As the government regulation forbids packets with weights lesser than what is specified on the packets, Mr. X has set the machine at a higher value, so that only packets with weights exceeding 1kg will be produced. This has created a problem for him. He feels that currently he is losing a lot of money in the way of excess material being packed. He has got an option to go for a more sophisticated packing machine at a certain cost that will reduce the variability. He wants to find out whether it is worthwhile going for the new machine. Say, the new process will produce packets with weight ranging from 1 to 1.05 kg, if set in the same manner.

A knowledge of p.d.f. of the weights produced by the current process will help Mr. X to calculate the probability that any packet will weigh more than, say, 1.05 kg. , or that any packet will weigh between 1.01 to 1.05 kg. These probabilities are helpful in his decision. A high probability of the weight exceeding 1.05 kg is an indicator of a high percentage of packets having more than 1.05 kg weight. These probabilities may help him calculate the expected loss due to the current process. This expected loss may be traded off then with the cost of buying the machine to arrive at the final decision.

**Situation 2**

Mr. T, a manufacturer of Electric bulbs, feels that the desired life of a bulb should be 100 hrs. , i.e. a new bulb should bum for 100 hrs. before the filament breaks. He realises that a high cost is associated with having a process that will manufacture all bulbs with life of more than 100 hrs. He is ready to make a trade off between the quality level and the cost.

In this case, if he knows the p.d.fs. of "the life (in hours)" of bulbs manufactured through different processes, then for different processes he can find out the probabilities that the life will exceed or equal 100 hrs. Suppose, he found the following for two processes:-

$\mathrm{P}(\text{ Iife } \leq 100\text{hrs.}) = .8$ for process 1

$\mathrm{P}(\text{ life } \geq 100\text{hrs.}) = .9$ for process 2

The above formula indicates that the process 2 is a better process, so far as quality is concerned. One may note that the cost for process 2 is higher than that of process 1. Mr X may now try to decide whether it is worthwhile paying extra cost for this quality.

The above formula shows how the information on p.d.f. can be helpful in decision making. This brings us to the question of assessing a p.d.f. Like we have seen in the cast-of discrete variables, for continuous variables also may real life situations can be approximated by certain theoretical distribution functions. Knowledge about the process of interest, and the past data, on the variable help us to find out what type of standard (theoretical) p.d.f. is to be applied in a particular situation.

We now present two important theoretical probability density functions, viz., the Exponential and the Normal. A study of the properties of these functions will be helpful in characterising the probability distributions in a variety of situations.

### Exponential Distribution

Time between breakdown of machines, duration of telephone calls, life of an electric bulb are examples of situations where the Exponential distribution has been found useful. In the previous unit, while discussing the discrete probability distributions, we have examined the Poisson process and the resulting Poisson distribution. In the Poisson process, we were interested in the random variable of number of occurrences of an event within a specific time or space. Thus, using the knowledge of Poisson process, we have calculated the probability that 0, 1, 2 …. accidents will occur in any month. Quite often, another type of random variable assumes importance in the context of a Poisson process. We may be interested in the random variable of the lapse of time before the first occurrence of the event. Thus, for a machine, we note that the first failure or breakdown of the machine may occur after 1 month or 1.5 months etc. The random variable of the number of failures within a specific time, as we have already seen, is discrete and follows the Poisson distribution. The variable, time of first failure, is continuous and the Exponential p.d.f. characterises the uncertainty.

If any situation is found to satisfy the conditions of a Poisson process, and if the average occurrence of the event of interest is m per unit time, then the number of occurrences in a given length of time t has a Poisson distribution with parameter mt, and the time between any two consecutive occurrences will be Exponential with parameter m. This can be used to derive the p.d.f. of the Exponential distribution.

Let f(t) denote the p.d.f. of the time between occurrence of the event F(t) denote the c.d.f. of the time between occurrence of the event (say, t >0).

Let A be the event that time between occurrence is less than or equal to t. and B be the event that time between occurrence is greater than t.

By definition, as A and B are mutually exclusive and collectively exhaustive : P(A) + P(B) = 1 ……… (1)

From the definition of c.d.f. and the description of event A, P(A) = F(t) ……….. (2)

From the definition of event B, as the time between occurrence is greater than t, it implies that the number of occurrences in the interval (0, t) is zero. Taking the distribution of number of occurrences in time t as Poisson, we can write':

P(B) = Probability that zero occurrences are there in time t, given that the average number of occurrences are mt.

From Poisson formula, P(B) can be written as :

P(B) $= \frac{e^{-mt} \times (mt)^0}{0!} = e^{-mt}$…….. (3)

From (1), (2) and (3), we have :

$F(t) + e^{-mt} = 1$

or, $\quad \mathrm{F(t)} = 1 - \mathrm{e}^{-\mathrm{mt}}$

Differentiating, we arrive at the p.d.f. :

$$f(t) = \begin{cases} m\mathrm{e}^{-\mathrm{mt}}, \mathrm{t} > 0 \\ 0, \text{otherwise.} \end{cases}$$

The above formula gives the pdf of the Exponential Distribution. We can now verify as to whether this is a valid pdf.

We find f(t) $\geq$ for all t as m>0

also $\int_0^\infty \mathrm{f(t)dt} = \int_0^\infty \mathrm{me}^{-\mathrm{mt}}\mathrm{dt} = 1$

Hence this is a valid p.d.f.

If we assume that the occurrence of an event corresponds to customers arriving for servicing, then the time between the occurrence would correspond to the inter-arrival time (IAT), and m would correspond to the arrival rate. Exponential has been used widely to characterise the IAT distribution. The Exponential p.d.f. is also used for characterising service time distributions. The parameter 'm' in that case, corresponds to the service rate. We take up an example to show the probability calculations using the Exponential p.d.f. In the final section of this unit, we will be illustrating through an example, the use of the Exponential distribution in decision-making.

**Example 5**

A highway petrol pump can serve on an average 15 cars per hour. What is the probability that for a particular car, the time taken will be less than 3 minutes?

**Solution**

Here, Exponential applies with m = 15 (service rate). We are interested in finding the probability that t <3 minutes i.e.t $< \frac{3}{60}$ hrs

From definition of c.d.f., we want to Find $\mathrm{F}\left(\frac{3}{60}\right) = \mathrm{F}\left(\frac{1}{20}\right)$

we have seen that $\mathrm{F(t)} = 1 - \mathrm{e}^{-\mathrm{mt}}$

$$\mathrm{F}\left(\frac{1}{20}\right) = 1 - \mathrm{e}^{-15\times1/20} = 1 - \mathrm{e}^{-3/4} = .5276$$

**Example 6**

The distribution of the total time a light bulb will burn from the moment it is first put into service is known to be exponential with mean time between failure of the bulbs equal to 1000 hrs. What is the probability that a bulb will burn more than 1000 hrs.

**Solution**

Here. $\mathrm{m} = \frac{1}{1000}$

and $f(t) = \begin{cases} \frac{1}{1000} e^{-t/1000} \text{ for } t \geqslant 0 \\ \quad 0 \text{ otherwise} \end{cases}$

We are interested in finding the probability that t >1000 hrs.

$$P(t > 1000) = 1 - p(t \leqslant 1000) = 1 - F(1000)$$

$$F(1000) = 1 - e^{-1000\times1/1000} = 1 - e^{-1}$$

∴ The required probability $= e^{-1} = 0.368$.

**Activity D**

In Example 5, find the probability that for any car, the time taken to service will be more than 10 minutes. Discuss how this probability and the probability you have found in Example 5, can be useful for the petrol pump owner.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

**Activity E**

In Example 6, find the probability that the life of any bulb will lie between 100 hrs. and 120 hrs. Elaborate as to how this information may be useful to the manufacturer of the bulb.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## Normal Distribution

The Normal Distribution is the most versatile of all the continuous probability distributions. It is found to be useful in Statistical inferences, in characterising uncertainties in many real-life processes, and in approximating other probability distributions.

Quite often, we face the problem of making inferences about processes based on limited data. Limited data is basically a sample from the full body of data on the process. Irrespective of how the full body of data is distributed, it has been found that the Normal Distribution can be used to characterise the sampling distribution. This helps considerably in Statistical Inferences.

Heights, weight and dimensions of a product are some of the continuous random variables which are found to be normally distributed. This knowledge helps us in calculating the probabilities of different events in varied situations, which in turn is useful for decision-making.

Finally, the Normal Distribution can be used to approximate certain probability distributions. This helps considerably in simplifying the probability calculations.

In the next few paragraphs we examine the properties of the Normal Distribution, and explain the method of calculating the probabilities of different events using the distribution. We then show the Normal approximation to Binomial distribution to illustrate how the probability calculations are simplified by using the approximation. An application of the Normal Distribution in decision-making is presented in the last section of the unit. The use of this distribution is Statistical Inferences is taken up in a later Block.

### Properties of the Normal Distribution

The p.d.f. of the Normal Distribution is given by :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}-\left(\frac{x-\mu}{\sigma}\right)^2} - \infty < x < \infty \ldots\ldots\ldots\ldots (1)$$

where, $\pi$ and e are two constants with values 3.14 and 2.718 respectively. The $\mu$ and $\sigma$ are the two parameters of the distribution, and x is a real number denoting the continuous random variable of interest.

The c.d.f. is given by:

$$F(X) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

It is apparent from the above that f is a positive function, $e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ being positive for any real number x. It can be shown that

$\int_{-\infty}^{+\infty} f(x)dx = 1$, so that f(x) is a valid p.d.f.

The mean and the standard deviation are respectively denoted by $\mu$ and $\sigma$. Thus, different values of these two parameters lead to different 'nominal curves'

The inherent similarity in all the normal curves' can be seen by examining the 'Standardised curve. The Standard Curve with $\mu = 0$ and $\sigma = 1$ is obtained by

using $Z = \frac{x-\mu}{\sigma}$, so that we get the p.d.f.

$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} - \infty < x < \infty$…………….. (2)

The p.d.f. (1) is referred to as the regular form, while the p.d.f. (2) is known as the standard form. Normal Distribution with mean $\mu$ and standard deviation σ is generally denoted by $\mathrm{N}(\mu, \sigma)$.

For large value of n, it is possible to derive the above p.d.f. as an approximation to the Binomial Distribution. The p.d.f. cannot be integrated analytically. The c.d.f. is tabulated for N(0,1) and the probabilities are calculate with the help of this table.

The plot of f(x) vs. x gives the Normal curve, and the area under the curve gives the probability. The Normal Distribution is symmetric about the mean; the area on each side of the mean is 0.5. The area between $\mu + K_1\sigma$ and $\mu + \mathrm{K}_2\sigma$ is the same for all Normal curves irrespective of the values of $\mu$and $\sigma$.

Though the range of the variable is specified from $-\infty$to$\infty$, 99.7% of the values of the random variable fall within $\pm 3\sigma$ limits, that is, $\mathrm{P}(\mu - 3\sigma \leq \mathrm{x} \leq \mu + 3\sigma) = .997$. Moreover, it is known that 95.4% and 68.6% of the values of the random variable lie between $\pm 2\sigma$and $\pm 1\sigma$ limits respectively.

Because of the symmetry, and the points of inflexion at $\pm 1\sigma$ distance, the Normal curve has a bell shape. The right and left tails of the curve extend indefinitely without touching the horizontal line.

### Probability Calculation

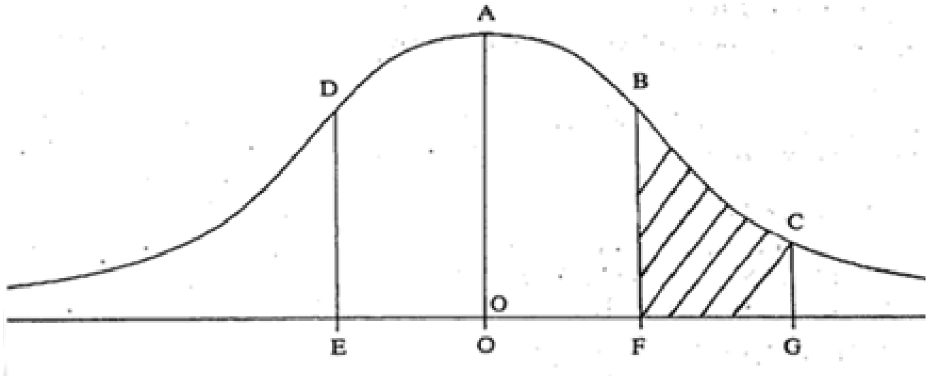
Suppose, it has been found that the duration of a particular project is normally distributed with a mean of 12 days and a standard deviation of 3. We are interested in finding the probability that the project will be completed in 15 days..

Given the $\mu$and $\sigma$ of the random variable of interest, we first find

$Z = \frac{x-\mu}{\sigma}$ Hence, $\mu = 12, \sigma = 3$ and $x = 15, \therefore Z = \frac{15-12}{3} = 1$

The values of the probabilities corresponding to Z are tabulated and can be found from the table. The Standard Normal being a symmetrical distribution, the table for one half (the right half) of the curve is sufficient for our purpose. The table gives the probability of Z being less than equal to a particular value.

Consider the following diagram depicting the Standardised Normal curve, denoted by N(0,1). The probability of Z lying between 1 and 2 can be represented by the area under the curve between Z values of 1 and 2; that is, the area represented by FBCG in the diagram given below.



Because of the symmetry, the area on the right of OA = area on the left of OA = 0.5. If you now look up a 'normal table' in any basic Statistics text book, you will find that corresponding to Z = 1.0, the probability is given as 0.3413. This only implies that the area OABF = 0.3413, so that,

$P(Z \leq 1) = 0.5 + 0.3413 = 0.8413$, the area to the left of OA being 0.5.

Similarly, corresponding to Z = 2.0, we find the value 0.4772 (area OACG = 0.4772). This implies,

$P(Z \leq 2) = 0.5 + 0.4772 = 0.9772$

∴ If we are interested in the shaded area FBCG, we find that, FBCG = Area OACG - Area OABF = 0.4772 - 0.3413 = 0.1359.

$\therefore P(1 \leq Z \leq 2) = 0.1359$.

The area, hence the probability, corresponding to a negative value of Z can be found from symmetry. Thus, we have the area OADE = the area OABF = 0.3413.

.: P(Z < -1) = 0.5 - 0.3413 = 0.1587.

Returning to our example, we are interested in finding the probability that the project duration is less than or equal to 15 days. Denoting the random variable by T, we know that T is N(12, 3).

$$\therefore P(T \leqslant 15) = P\left(\frac{T-12}{3} \leqslant \frac{15-12}{3}\right) = P(Z \leqslant 1) = 0.5 + 0.3413 = 0.8413$$

Similarly, if we were interested in finding out the chances that the project duration will be between 9 and 15 days, we can proceed in a similar way.

$$\therefore P(9 \leq T \leq 15) = P\left(\frac{9-12}{3} \leq \frac{T-12}{3} \leq \frac{15-12}{3}\right) =$$

$$P(-1 \leqslant Z \leqslant 1) = 0.3413 + 0.3413 = 0.6813$$

(Note that this confirms our earlier statement that 68% of the values lie between $\pm 1\sigma$ limit.)

**Normal as an Approximation to Binomial**

For large n and with p value around 0.5, the Normal is a good approximation for the Binomial. The corresponding μ, and σ for the Normal are np and $\sqrt{npq}$ respectively. Suppose, we want to find the probability that the number of heads in a toss of 12 coins will lie between 6 and 9. From the previous unit, we know that this probability is equal to :

$$\sum_{x=6}^{9} \ {}^{12}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{12-r}$$

As such, this tedious calculation can be obviated by assuming that the random variable, number of heads (H), is Normal with mean = np and $\sigma = \sqrt{\text{npq}}$. Here $\mu = 12 \times 0.5 = 6$ and $\sigma = \sqrt{12 \times 0.5 \times 0.5} = \sqrt{3} = 1.732$

∵ assuming H is N (6, 1.732), we can find the probability that H lies between 6 and 9. The following continuity correction helps in better approximation. Instead of looking for the area under the Normal curve between 6 and 9, we look up the area between 5.5 and 9.5, i.e. 0.5 is included on either side.

$$\therefore P(5.5 \leqslant H \leqslant 9.5) = P\left(\frac{5.5-6}{1.732} \leq \frac{H-6}{1.732} \leq \frac{9.5-6}{1.732}\right)$$

$$\therefore P(-.289 \leqslant Z \leqslant 2.02)$$

From the table, corresponding to Z = 0.289 and 2.02 we find the values 0.114 and 0.4783.

∴ the required probability = 0.114 + 0.4783 = 0.5923, Now you may check that by using the Binomial distribution, the same probability can be calculated as 0.5934.

### Fractile of a Normal Distribution

The concept of Fractile as applied to Normal Distribution is often found to be useful. The kth fractile of $N(\mu, \sigma)$ can be found as follows. First we find the $k^{th}$ fractile of the N(0,1). Let $Z_k$ be the $K^{th}$ fractile of N(0,1).

By definition, $F(Z_k) = K$, $(0 < K < 1)$.

Say, if $Z_k$ is the .975th fractile of N(0,1), then

$$F(Z_k) = 0.975, P(Z \leq Z_k) = 0.975 = 0.5 + 0.475$$

From the table, we find that corresponding to Z = 1.96, the probability is 0.475. Hence $Z_k$ = 1.96. Now suppose that we are interested in the 0.975th fractile of N(50,6). If $X_k$ be the required fractile,

then $\frac{X_k - \mu}{\sigma} = Z_k$

$$\therefore X_k = \mu + Z_k\sigma = 50 + 1.96 \times 6 = 61.76$$

From symmetry, the .025th fractile of N(50,6) can be seen to be = 50 - 1.96 x 6 = 38.24.

### Activity F

A ball-bearing is manufactured with a mean diameter of 0.5 inch and a standard deviation in diameters of .002 inch. The distribution of the diameter can be considered to be normal. The bearing with less than .498 inch and more than .0502 inch are considered to be defective. What is the probability that a ball - bearing manufactured through this process will be defective ?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity G**

Suppose from the above exercise, you have found that the probability of a defective is 0.32. If the bearing are packed in lots of 100 units and sent to the supplier, what is the probability that in any such lot, the number of defectives will be less than 27? (The probability corresponding to Z value of 1.07 is 0.358.)

## 7.4 APPLICATIONS OF CONTINUOUS DISTRIBUTIONS

The following two examples illustrate the use of the Exponential and the Normal Distribution in decision-making.

**Example 7**

A TV manufacturer is facing the problem of selecting a supplier of Cathode-ray tube which is the most vital component of a TV. Three foreign suppliers, all equally dependable, have agreed to supply the tubes. The price per tube and the expected life of a tube for the three suppliers are as follows :

| | Price/tube | Expected life per tube |
|---|---|---|
| Supplier 1 | Rs. 800 | 1500 hrs. |
| Supplier 2 | Rs. 1000 | 2000 hrs. |
| Supplier 3 | Rs. 1500 | 4000 hrs. |

The manufacturer guarantees its customers that it will replace the TV set if the tube fails earlier than 1000 hrs. Such a replacement will cost him Rs. 1000 per tube, over and above the price of the tube.

Can you help the manufacturer to select a supplier?

**Solution**

The Expected cost per tube for each supplier can be found as follows :

Expected cost per tube = price per tube + expected replacement cost per tube.

Expected replacement cost per tube is given by the product of the cost of replacement and the probability that a replacement is needed. Both the cost of replacement and the probability vary from supplier to supplier. We note that, a replacement is called for if the tube fails before 1000 hrs., so that, for each supplier we can calculate the P(life of tube $\leq$ 1000 hrs.). This probability can be calculated by assuming that the time between failure is exponential. Thus, $p(t \leq 1000)$ is basically exponential with

$m = \frac{1}{1500}, \frac{1}{2000}$, and $\frac{1}{4000}$, for the three suppliers

$$p\left(t \leqslant 1000/m = \frac{1}{1500}\right) = F_1(1000) = 1 - e^{-1000/1500} = .4866$$

$$\mathbf{p}\left(t \leqslant 1000/m = \frac{1}{2000}\right) = F_2(1000) = 1 - e^{-1000/2000} = .3935$$

and $p\left(t \leqslant 1000/m = \frac{1}{4000}\right) = F_3(1000) = 1 - e^{-1000/4000} = .2212$

Once the expected costs for each supplier are known, we can take a decision based on the cost. The calculations are shown in the table below :

| **Supplier Number** | **Price per tube P** | **Cost per Replacement C** | **P(life 1000 hrs.) P** | **Expected cost per tube E=(P+Cp)** |
|---|---|---|---|---|
| 1 | 800 | 1800 | .4886 | 1679.48 |
| 2 | 1000 | 2000 | .3935 | 1787 |
| 3 | 1500 | 2500 | .2212 | 2053 |

We find that for the supplier 1, the expected cost per tube is the minimum. Hence the decision is to select 1.

**Example 8**

A supplier of machined parts has got an order to supply piston rods to a big car manufacturer. The client has specified that the rod diameter should lie between 2.541 and 2.548 cms. Accordingly, the supplier has been looking for the right kind of machine. He has identified two machines, both of which can produce a mean diameter of 2.545 cms. Like any other machine, these machines are also not perfect. The standard deviations of the diameters produced from the machine 1 and 2 are 0.003 and 0.005 cm. respectively, i.e. machine 1 is better than machine 2. This is reflected in the prices of the machines, and machine 1 costs Rs. 3.3 lakhs more than machine 2. The supplier is confident of making a profit of Rs. 100 per piston rod; however, a rod rejected will mean a loss of Rs. 40.

The supplier wants to know whether he should go for the better machine at an extra cost.

**Solution**

Assuming that the diameters of the piston rods produced by the machining process is normally distributed, we can find the probability of acceptance of a part if produced in a particular machine.

For machine 1, we find that the diameter is N(2.545,.003), and for machine 2, we find that the diameter is N(2.545,.005)

If D denote the diameter, then :

$2.541 \leq D \leq 2.548$, implies the rod is accepted.

Probability of acceptance if a rod is produced in machine 1

$= P(2.541 \leqslant 2.548)$

$= P\left(\frac{2.541-2.545}{.003} \leqslant Z \leqslant \frac{2.548-2.545}{.003}\right)$

$= P(-1.33 \leqslant Z \leqslant 1)$

$= .4066 + .3413 = .7479$[ from $N(0,1)$ table ]

Hence probability of rejection = 1 - .7479 = .2521

Expected profit per rod if machine 1 is used

= 100 x .7479 - 40 x .2521 = Rs. 64.706 (1)

Similarly, if machine 2 is used, we can find the expected profit per rod Probability of acceptance here

$= p\left(\frac{2.541-2.545}{.005} \leq Z \leq \frac{2.548-2.545}{.005}\right)$

$= p(-.8 \leq D \leq .6)$

=.2881+.2257 = .5138

Probability of rejection = 1 - .5138 = .4862

Expected profit per rod if machine 2 is used

= 100 x .5138 - 40 x .4862 = Rs. 31.932 (2)

Thus, from (1) and (2), we find that the expected profit per part is more if machine 1 is used. As machine 1 costs 3.3 lakh more than machine 2, it will be profitable to use machine 1 only if the production is more.

We can find the breakeven production level as follows.

Let N be the number of rods produced, for which both the machines are equally profitable.

Then N x (64.706 - 31.932) = 3,30,000

or; N = 10,069

This implies that it is advisable to go in for machine 1, only if the production level is higher than 10,070. (Note that we assume that there is enough demand for the rods.)

**Activity. H**

Suppose in Example 8, you have decided that machine 1 should be used for production. Assume now, that this machine has got a facility by which one can set the mean diameter, i.e., one can set the machine to produce any one mean diameter ranging from 2.500 to 2.570 cm. Once the machine is set to a particular value, the rods are produced with mean diameter equal to that value and standard deviation equal to 0.003 cm. If the profit per rod and loss per rejection is same as in example 8, what is the optimal machine setting?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

## 7.5 SUMMARY

The function that specifies the probability distribution of a continuous random variable is called the probability density function (p.d.f.). The cumulative density function (c.d.f.) is found by integrating the p.d.f. from the lowest value in the range upto an arbitrary level x. As a continuous random variable can take innumerable values in a specified interval on a real line, the probabilities are expressed for interval rather than for individual values. In this unit, we have examined the basic concepts and assumptions involved in the treatment of continuous probability distributions. Two such important distributions, viz., the Exponential and the Normal have been presented. Exponential distribution is found to be useful for characterising uncertainty in machine life, length of telephone call etc., while dimensions of machined parts, heights, weights etc. found to be Normally distributed. We have examined the properties of these p.d.fs. and have seen how probability calculations can be done for these distributions. In the final section, two examples are presented to illustrate the use of these distributions in decision-making.

## 7.6 FURTHER READINGS

Chance, W., Statistical Methods for Decision Making, R. Irwin Inc. Homewood.

Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley & Sons Inc.: New York.

Gangolli, R.A. and D. Ylvisaker. Discrete Probability, Harcourt Brace and World. Inc.: New York.

Levin, R., Statistics for Management, Prentice-Hall Inc., New York.

Parzen, E., Modern Probability Theory and Its Applications, Wiley: New York.

Nick T. Thomopoulos. Probability distribution, Springer

# UNIT 8 DECISION THEORY

**Objectives**

After reading this unit, you should be able to:

- structure a decision problem involving various alternatives and uncertainties in outcomes
- apply marginal analysis for solving decision problems under uncertainty
- analyse sequential problems using Decision Tree Approach
- appreciate the use of Preference Theory in decision-making under uncertainty
- analyse uncertain situations where probabilities of outcomes are not known.

**Structure**

8.1 Introduction
8.2 Certain Key Issues in Decision Theory
8.3 Marginal Analysis
8.4 Decision Tree Approach
8.5 Preference Theory
8.6 Other Approaches
8.7 Summary
8.8 Further Readings

ignou
THE PEOPLE'S
UNIVERSITY

## 8.1 INTRODUCTION

In every sphere of our life we need to take various kinds of decisions. The ubiquity of decision problems, together with the need to make good decisions, have led many people from different time and fields, to analyse the decision-making process. A growing body of literature on Decision Analysis is thus found today. The analysis varies with the nature of the decision problem, so that any classification base for decision problems provides us with a means to segregate the Decision Analysis literature. A necessary condition for the existence of a decision problem is the presence of alternative ways of action. Each action leads to a consequence through a possible set of outcome, the information on which might be known or unknown. One of the several ways of classifying decision problems has been based on this knowledge about the information on outcomes. Broadly, two classifications result:

a) The information on outcomes are deterministic and are known with certainty, and

b) The information on outcomes are probabilistic, with the probabilities known or unknown.

The former may be classified as Decision Making under certainty, while the latter is called Decision Making under uncertainty. The theory that has resulted from analysing decision problems in uncertain situations is commonly referred to as Decision Theory. With our background in the Probability Theory, we are in a position to undertake a study of Decision Theory in this unit. The objective of this unit is to study certain methods for solving decision problems under uncertainty. The methods are consequent to certain key issues of such problems. Accordingly, in the next section we discuss the issues and in subsequent sections we present the different methods for resolving them.

## 8.2 CERTAIN KEY ISSUES IN DECISION THEORY

Different issues arise while analysing decision problems under uncertain conditions of outcomes. Firstly, decisions we take can be viewed either as independent decisions, or as decisions figuring in the whole sequence of decisions that are taken over a period of time. Thus, depending on the planning horizon under consideration, as also the nature of decisions, we have either a single stage decision problem, or a sequential decision problem. In real life, the decision maker provides the common thread, and perhaps all his decisions, past, present and future, can be considered to be sequential. The problem becomes combinatorial, and hence difficult to solve. Fortunately, valid assumptions in most of the cases help to reduce the number of stages, and make the problem tractable. In earlier unit, we have seen a method of handling a single stage decision problem. The problem was essentially to find the number of newspaper copies the newspaper man should stock in the face of uncertain demand, such that, the expected profit is maximised. A critical examination of the method tells us that the calculation becomes tedious as the number of values the demand is taking increases. You may try the method with a discrete distribution of demand, where demand can take values from 31 to 50. Obviously a separate method is called for. We will be presenting Marginal Analysis for solving such single stage problems. For sequential decision problems, the Decision Tree Approach is helpful and will be dealt with in a later section. The second issue arises in terms of selecting a criterion for deciding on the above situations. Recall as to how we have used 'Expected Profit' as a criterion for our decision. In both the Marginal Analysis and the Decision Tree Approach, we will be using the same criterion. However, this criterion suffers from two problems. Expected Profit or Expected Monetary Value (EMV), as it is more commonly known, does not take into account the decision maker's attitude towards risk. Preference Theory provides us with the remedy in this context by enabling us to incorporate risk in the same set up. The other problem with Expected Monetary Value is that it can be applied only when the probabilities of outcomes are known. For problems, where the probabilities are unknown, one way out is to assign equal probabilities to the outcomes, and then use EMV for decision-making. However this is not always rational, and as we will find, other criteria are available for deciding on such situations.

For the purpose of this unit, we will be discussing the issues as raised above. This will be achieved through a study of the following:

1) Marginal Analysis for single stage decision problems.
2) Decision Tree Approach for sequential decision problems.
3) Preference Theory.
4) Other approaches for problems where probabilities are unknown.

In the subsequent sections we take up the above in the order presented.

**Activity A**

Suppose you have the option of investing either in Project A or in Project B. The outcomes of both the projects are uncertain. If you invest in Project A, there is a 99' chance of making Rs. 20,000 profit, and a 1% chance of losing Rs. 1,00,000. If project B is choosen, there is a 50-50 chance of making a profit of Rs. 6,000 or Rs. 18,000. Which project will you choose and why?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity B**

Suppose in Exercise 1, you have calculated the expected projects as follows.

$$\mathrm{EMV_A} = 99 \times 20{,}000 - .01 \times 1{,}00{,}000 = \mathrm{Rs.}\,18{,}000$$

$$\mathrm{EMVB} = .5\mathrm{X}6{,}000 - .5 \times 18{,}000 = \mathrm{Rs.}\,12{,}000$$

You have thus found that by investing in Project A, you can expect more money, so you have chosen A. Your friend, when given the same option, chooses B, arguing that he would not like to go bankrupt (losing 1 lakh) by choosing A. How do you reconcile these two arguments?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 8.3 MARGINAL ANALYSIS

We have seen how expected value can be used while deciding on one alternative from among several alternative courses of actions, each of which is characterised by a set of uncertain outcomes. It is easy to see that the computations become tedious as the number of values, the random variable can take, increases. Consider the example of the newspaper man discussed in section 6.4. Instead of six values of the demand that we have assumed there, if the demand could take, say, twenty values, with different chances of occurrence of each Value, the computation would become very tedious. In such cases, marginal analysis is very helpful. In this section, we explain the concept behind this analysis.

Consider Example 1 in section 6.4 with the following change. Let us assume that the newspaper man has found from the past data that the demand can take values ranging from 31, 32... to 50. For easy representation, let us assume that each of these values has got an equal chance of occurrence, viz. ,$\frac{1}{20}$. The problem is to decide on the number of copies to be ordered.

Marginal Analysis proceeds by examining whether ordering an additional unit is worthwhile or not. Thus, we will order X copies, provided ordering the Xth copy is worthwhile but ordering the (X+1)th copy is not. To find out whether ordering X copies is worthwhile, we note the following. Ordering of the Xth copy may meet with two consequences, depending on the occurrences of two events:

A The copy can be sold.

B The copy cannot be sold.

The Xth copy can be sold only if the demand exceeds or equals X, whereas, the copy cannot be sold if the demand turns out to be less than X. Also, if event A occurs, we will make a profit of 50 p. on the extra copy, and if even B occurs, there will be a loss of 30 p. As this profit and loss pertains to the additional or marginal unit, these are referred to as marginal profit or loss and the resulting analysis is called marginal analysis.

Using the following notations

$$\mathrm{K}_1 = \text{ Marginal profit} = 50\mathrm{p}$$

$$\mathrm{K}_2 = \text{ Marginal loss} = 30p$$

$$\mathrm{P(A)} = \text{ Probability ( Demand } \geq \mathrm{X}) = 1 \text{ -Probability (Demand } \leq \mathrm{X} - 1 \text{ ).}$$

$$\mathrm{P(B)} = \text{ Probability ( Demand } < \mathrm{X}) = \text{ Probability (Demand } \leq \mathrm{X} - 1 \text{ ).}$$

We can write down the expected marginal profit and expected marginal loss as

$$\text{Expected Marginal Profit} = \mathrm{K}_1\mathrm{P(A)}$$

$$\text{Expected Marginal Loss } = \mathrm{K}_2\mathrm{P(B)}$$

Ordering the Xth copy is worthwhile only if the expected profit due to it is more than the expected loss, so that

$$\mathrm{K}_1\mathrm{P(A)} \geq \mathrm{K}_2\mathrm{P(B)}$$

Now, if F(D) denotes the c.d.f. of demand, then by definition, Probability

$$\text{Demand } \leq (\mathrm{X} - 1) \pm \mathrm{F}(\mathrm{X} - 1)$$

Hence, $K_1[1 - F(X-1)] \geq K_2F(X-1)$

or; $K_1 - K_1F(X-1) - K_2F(X-1) \geq 0$

or; $F(X-1) \leq \frac{K_1}{K_1+K_2}$ …. (CONDITION 1)

Thus, if condition 1 holds good, it is worthwhile to order the Xth copy.

If the optimal decision is to order X copies, then ordering the (X+1)th copy will not be worthwhile, i.e. the expected marginal profit due to the (X+1)th copy should be less than the expected loss.

Proceeding with the analysis in the same way as above, we have :

$$\text{Expected Marginal Profit} = K_1 \text{ Probability (Demand} \geq X+1)$$

$$= K_1[1 - F(X)]$$

Expected Marginal Loss $= K_2\, F(X)$

∴ For the (X+1)th copy : $K_1[1 - F(X)] \leq K_2F(X)$

From conditions (1) and (2) and the definition of Fractile, it is clear that X will be the $\left(\frac{K_1}{K_1+K_2}\right)^{th}$ the fractile of the Demand distribution.

Thus, for our problem, given the above result, all that we have to do is to calculate $K = \frac{K_1}{K_1+K_2}$ and find the Kth fractile of the distribution, which will give us the required answer.

In our problem :

$K = \frac{.5}{.5+.3} = .625$ and the .625th fractile is 43.

∴ The optimal decision is to order 13 copies.

We can verify quickly that in the problem given in section 10.4, the .625th fractile of the demand distribution is 33. So the optimal decision there is to order 33, which is the answer that we have obtained there.

The above shows how marginal analysis helps us in arriving at the optimal decision with very little computation. This is especially useful when the random variable of interest takes a large number of values. Though we have demonstrated this for a discrete demand distribution the same logic can be shown to be applicable for continuous distributions also. Instead of the distribution we have taken, if we would have assumed that demand is normal with a specific $\mu$ and $\sigma$, then also the same $K^{th}$ fractile of $N(\mu, \sigma)$ would have given us the optimal decision.

**Activity C**

The demand for a particular perishable item is known to be N (50, 6). The cost of under stocking $(K_1)$, and the cost of overstocking $(K_2)$ per unit is known to be Rs. 20 and Re. 1 respectively. How much of the item should be stocked to minimise the cost due to understocking and overstocking?

(Note that understocking implies stocking less than what is demanded, the loss being in terms of contribution, while overstocking implies stocking more than what is demanded, and hence, there is the cost of not being able to sell. These are $K_1$ and $K_2$ respectively as discussed in the text.)

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 8.4 DECISION TREE APPROACH

In the earlier section we have seen a single stage decision problem. Quite often the decision maker has to take decisions in a sequence, the decisions coming later in the sequence being dependent on those coming earlier. The sequence is either built-in, or it is possible to engineer such a sequence for a better decision. For example, consider the periodic production decision for a certain item with uncertain demand (say, refrigerator); for each period, a decision on the number of units to be produced is to be taken, given the uncertainties in demand during different periods. Thus, we will have a number of decisions for each period, with intervening uncertainties in outcomes for each decision between any two periods. In such cases, the sequence is built-in.

In contrast to the above, we find situations, where the time-frame of decisions are such, that before going for the final decision, it is possible to go for a method for generating extra information that will facilitate the final decision, For example, before deciding on marketing a product nationally, one can decide on Test Marketing. Similarly, in a production situation, where a machine produces an unknown percentage of defectives, one may have an option to buy a special attachment that helps to produce a known low fraction of defectives. The trade-off then, is between not buying the attachment and thereby risking a high percentage of defectives, of buying the attachment at a cost, to safeguard against the risk. An infinite sequence of decisions can be engineered in this case by allowing sampling from the current process, to ascertain the percentage of defectives. Thus, at each stage we can have two alternatives :

a) buying, and

b) not buying and sampling.

This can go on till we decide to stop sampling due to some reason (e.g. sampling cost becomes prohibitive).

The Decision Tree Approach provides us with a useful way to analyse such sequential decision problems. We illustrate this approach through an example. The oil drilling example has been a favourite of many authors.

## Example I

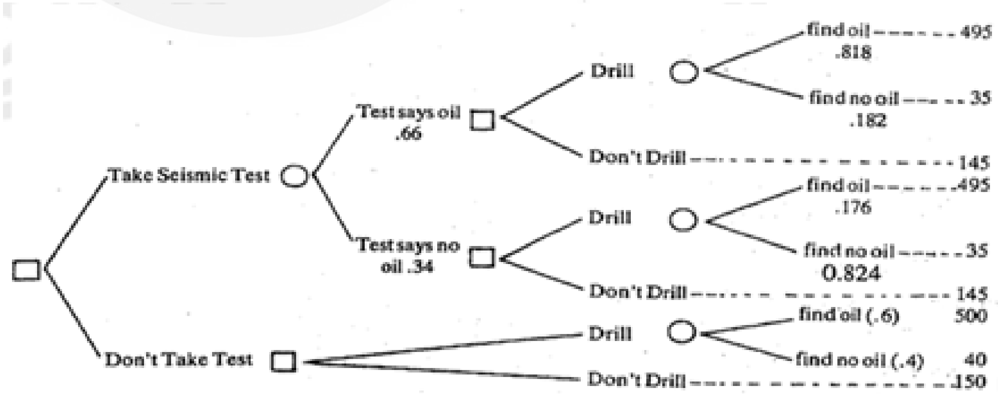
Consider the decision of drilling for oil in a particular region, confronting our decision maker. The chances of getting oil in the region as per the geologist's report is known to be 0.6. To start with, the decision maker has got Rs. 1.5 lakh. The consequences of drilling and getting oil and that of drilling and not getting oil, in terms of cash left after decision, are known to be Rs. 5 lakh and Rs. 40,000 respectively. The decision maker has got an option to undertake a seismic test that will increase his knowledge about the oil content of the region. The test will cost him Rs. 5,000; however, the benefit in having the test is that, if oil is actually there the test would predict it correctly for 90% of the time; and if there is actually no oil, that would be predicted correctly for 70% of the time. What should we do and why?

The first step is to structure the decision problem. In Decision Tree Approach a square “□” is used to denote an action or a decision point, and a circle “□” is used to, illustrate the point of uncertainty. First the alternatives of courses of action are shown as emanating from the decision point and then corresponding to each decision, the possible outcomes are shown emanating from the uncertainty point. The probability and consequence for each outcome are listed by the side of the outcome. The resulting diagram is called a Decision Tree. For our example, we have to start with two possible actions:

1) Take the Seismic Test
2) Do not take the Seismic Test

If the test is taken, the test may say that there will be oil, or it may say that there will not be any oil. These outcomes are uncertain as the test is not a perfect test. Once the test outcomes are known, the decision maker has again to decide on whether to drill or not. The outcomes corresponding to each decision are once again known here. Similarly, If it is decided that the test is not to be taken, one has to still decide on whether to drill or not.

The Decision Tree, thus, can be drawn as follows:



The sequences shown beside each outcome are in thousand rupees.

The second step is to write down the probabilities corresponding to each outcome. If the test is not taken, the chances of finding oil is given directly by the geologist's report as 0.6. Therefore, the chances of not getting oil = 1-.6 = .4. These can then be written corresponding to each of the outcomes with consequences of 500 and 40 thousand. However, once the test is taken, the chances of the test saying positive (presence of oil) or negative (no oil) is dependent on the predictive capability of the test, and has to be calculated. Similarly, the probability of finding oil given that test has yielded positive results is expected to be more than 0.6. These and related probabilities are to be calculated. also. The probability calculations can be done by using Bayes' Theorem.

Using the same notations, we find two mutually exclusive and collectively exhaustive events A and B as follows :

A : find oil

B : find no oil

The other events defined in the context of the same experiment are :

C : Test says oil is there (positive results).

D : Test says no oil is there (negative results).

The data given to us are

P(A) = Probability of finding oil = 0.6

P(B) = Probability of not finding oil = 0.4

P(C/A) = Probability test predicts correctly when oil is actually there = 0.9

P(D/A) = Probability test predicts incorrectly when oil is actually there = 0.1

P(D/B) = Probability test predicts correctly when actually oil is not there = 0.7

P(C/B) = Probability test predicts incorrectly when actually no oil is there = 0.3

We are interested in finding

P(C) = Probability that test says oil is there.

P(D) = Probability that test says no oil is there.

P(A/C) = Probability of finding oil, given positive test results.

P(A/D) = Probability of finding oil, given negative test results.

P(B/C) = Probability of finding oil, given positive test results.

P(B/D) = Probability of finding oil, given negative test results.

We have Bayes' Theorem:

$$\mathrm{P(A/C)} = \frac{\mathrm{P(C/A) \times P(A)}}{\mathrm{P(C/A)P(A) + P(C/B)P(B)}} = \frac{0.9 \times 0.6}{.9 \times .6 + .3 \times .4} = .818$$

$$P(B/C) = \frac{P(C/B) \times P(B)}{P(C/B)P(B) + P(C/A)P(A)} = \frac{0.3 \times .4}{.3 \times .4 + .9 \times .6} = .182$$

$$\mathbf{P}\mathrm{(A/D)} = \frac{\mathrm{P(D/A) \times P(A)}}{\mathrm{P(D/A)P(A) + P(D/B)P(B)}} = \frac{.1 \times .6}{.1 \times .6 + .7 \times .4} = .176$$

$$\mathbf{P}(\mathbf{B}/\mathrm{D}) = \frac{\mathrm{P(D/B)} \times \mathbf{P}\mathrm{(B)}}{\mathrm{P(D/A)}\mathbf{P}\mathrm{(A) + P(D/B)P(B)}} = \frac{.7 \times .4}{.1 \times .6 + .7 \times .4} = .824$$

ignou THE PEOPLE'S UNIVERSITY

We also know that,

$$P(C) = P(C/A)P(A) + P(C/B)P(B) = .9 \times .6 + .3 \times .4 = .66$$

$$P(D) = P(D/A)P(A) + P(D/B)P(B) = .1 \times .6 + .7 \times .4 = .34$$

$$[\text{ Check } P(C) + P(D) = 1, P(A/C) + P(B/C) = 1, P(A/D) + P(B/D) = 1]$$

These probabilities are incorporated in the decision tree diagram. The final step consists of finding the Expected Monetary Value (EMV) for the decisions. We start from the Northeast corner of the diagram and "fold back" the tree as follows :

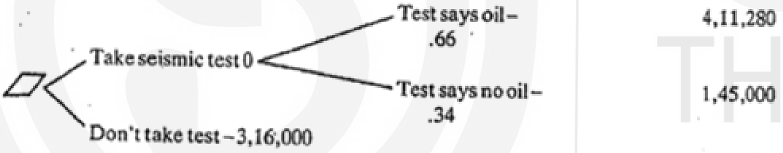
The extreme Northeast decision is "to drill" with the outcomes of finding oil. or not finding oil with chances of occurrence of .818 and .182. The respective contributions are Rs, 4,95,000 and Rs. 35,000.

∴ EMV of decision to drill = 4,95,000 × .818 + 35,000 × .182 = Rs. 4,11,280

This being greater than the payoff due to not drilling (1,45,000), we can say that once the test says oil, it is better to go for drilling, and the corresponding expected payoff in that case is Rs. 4,11,280.

Similarly, when the test says no oil, we find that "not drilling" is a better option than “drilling", as the expected payoff in the former is more (Rs. 1,45,000) vis-a-vis the latter (= .176 × 495 + .824 × 35 = 115960).

The earlier diagram is thus reduced as shown:



If test is not taken, the expected payoff of drilling is:

500 × .6 + 40 × .4 = 3,16,000

This being greater than not drilling (1,50,000) it is better to go for drilling if the test has not been taken. This is shown in the diagram. We now calculate the EMV of taking a seismic test :

.66 × 4,11,280 + .34 ×1,45,000 = 3,20,745

Therefore, as this payoff is more than what one can expect if the test is not taken, it is better to take the test.

Hence, the decision is to "Take the Test". If the test result says no oil then one should not drill, and if the test result is positive one should drill. This decision will maximise the EMV.

**Activity D**

ABC Company is a small time manufacturer of L.P. records. The record business is almost a monopoly of another Calcutta Based company (XYZ), and ABC's ability to. survive so far may be attributed to their able and experienced Managing Director Mr. A. As all the topmost artists are under

ignou THE PEOPLE'S UNIVERSITY

the contract of XYZ, ABC's strategy has been to get hold of new faces for recording. Mr. A's intuition in this respect has proved useful. He has been actively participating in recruiting new faces, and he believes that apriori 70% of his recruits stand the chance of being successful nationally. Once a new face is chosen, a tape is cut and an initial production of 5,000 records is undertaken for test marketing. It has been found that when the, recruit is actually a success nationally, test marketing would have predicted the outcome 90% of the time, and when the recruit is actually a failure nationally, the outcome would have been predicted 70% of the time. Based on test marketing results, the decision to go for national marketing is taken up. National marketing involves a production of 50,000 records. The artist is paid a sum of 5,000 once a tape in cut. The variable cost per record for production run of 5,000 and 50,000 works out to Rs. 13 per record and Rs. 10 per record respectively and the selling price is Rs. 40 per record.

Mr. A is thinking of entering the ghazal market, and has currently recruited a ghazal, singer, He feels that the prediction capability of test marketing will be on the lower side for ghazals: His estimate is that the test marketing would predict a success, when it is actually a success for only 70% of the time (as against 90% earlier), and in case of failure, it would predict correctly only 60% of the time (as against 70% earlier). Given the low prediction capability, he is wondering whether it is worthwhile to go for test marketing at all.

Can you help him in his decision? You may assume that a success in case of test or National marketing would imply an ability to sell 5,000 and 50,000 records respectively, whereas a failure in both cases would amount to zero sales, for all practical purposes.

## 8.5 PREFERENCE THEORY

So far, while deciding on an action, we have used the criterion of maximising the EMV or expected payoff. This does not take into account the decision maker's attitude towards risk. If a company is financially weak, it may decide not to use the EMV maximising action, if there is even a small chance of going bankrupt following that action. Preference Theory helps us in such situations by providing a systematic way of measuring the consequences on a preference scale, that reflects the decision maker's attitude towards risk. The objective of this section is to illustrate how Preference Theory can be used for decision-making.

The procedure consists of eliciting information from the decision maker (d.m.), on his 'certainly equivalents' (CE) corresponding to each alternative; CE of an alternative being the amount he is ready to exchange for the uncertain consequences of the particular alternative. For example, consider any alternative of investing in a project, the possible outcomes of which are (a) net loss of Rs. 1,00,000 with probability 0.1, and (b) net gain of Rs. 20,000 with probability 0.9. Now, if the d.m. is risk averse, he might not like even the small odds of losing 1 lakh, and he might be content in having an alternative paying him a certain amount of Rs. 5,000 as against the above (EMV of above Rs. 8,000). You can imagine that this investment gamble is the exclusive right of a class of people, and our d.m. is one among them. Thus, if this exclusive right is allowed to be sold to other people, the d .m. is ready to sell it for Rs. 5,000. The difference between the EMV and the CE is defined as the risk premium. Here, CE is Rs. 5,000; hence the risk premium is Rs. 3,000.

As the number of alternatives increase, it becomes difficult to collect preference information in this way. The Preference curve, which is a plot of the monetary value (X - axis) and the preference (Y- axis) is then obtained as follows. First, the best and the worst consequences corresponding to any decision are identified. The preference values of 1 and 0 are then given corresponding to the best and worst consequences respectively, giving us two points in the Preference curve. The step for obtaining the subsequent points are given below :

Let $R_0$ = Consequence corresponding to worst decision.

$P(R_0)$ = Preference corresponding to $R_0 = 0$.

$R_1 =$ Consequence corresponding to the best decision.

$P(R_1) =$ Preference corresponding to $R_1 = 1$

Step 1 We find the d.m 's CE of a 50-50 chance of getting Rs. $R_0$ or Rs. $R_1$. Suppose, he gives the value Rs. $(CE_1)$.

Step 2 We find the preference corresponding to $CE_1$ i.e. $P(CE_1)$.

Preference of an alternative is defined as the mathematical expectation of preferences corresponding to the consequences of the alternative. A preference P(x) assigned to a consequence x implies that the d.m. is indifferent to having an amount x for certain or having uncertain consequences of (a) [1-p(x)] of Rs . $R_0$ and (b) P(x) of achieving Rs. $R_I$.

$$\therefore P(CE_1) = .5 \times 0 + 0.5 \times 1 = .5$$

Step 3 Now, we ask the d.m., as to what certain amount would make him indifferent to uncertain consequences of Rs. $(CE_1)$ with probability 0.5 and Rs. $R_1$ with probability 0.5. Say, he says Rs. $(CE_2)$.

Step 4 We find $P(CE_2) = 0.5P(CE_1) + 0.5P(R_1) = .5 \times .5 + .5 \times 1 = .75$

Step 5 We continue till sufficient values of P(x) corresponding to different x are generated, and the curve of P(x) vs x can be drawn.

Once the preference curve is drawn, the preferences corresponding to each consequence of the problem can be obtained. In the same Decision Tree, the consequence can now be replaced by the preferences and the criterion of maximising expected preference be used for arriving at the decision. We now illustrate the above through an example.

**Example 2**

Let us take Example 1 of the earlier section. Suppose the decision maker is not a player of long run average (expected value). We want to get his preference curve for the problem, and arrive at the decision that maximises his expected preference.

**Solution**

We obtain the Preference curve of the d.m. as follows :

Step 1 From the Decision Tree of the earlier section, we see the worst consequence Rs. 35,000

the best consequence = Rs. 5,00,000

Question to d.m. : Suppose you have got a 50-50 chance of getting Rs, 35,000 or Rs. 5,00,000; for what certain amount will you exchange it?

Answer : Suppose he says Rs. 1,00,000 i.e.$CE_1$ = Rs. 1,00,000.

Step 2

Question to d.m.: Suppose you have a 50-50 chance of getting Rs. 1 lakh or Rs. 5 lakh, for what certain amount will you exchange it?

Answer : $CE_2$ = Rs. 2 lakh.

Step 3

Question to d.m.: What is your CE for a 50-50 chance of getting Rs. 2 lakh or Rs. 5 lakh.

Answer : $CE_3$ = Rs. 2.5 lakh.

Step 4. Continue questioning to obtain CE values till sufficient points, are there to draw a graph.

Step 5 Calculate $P_1, P_2$, P3 ……… the preference corresponding to $CE_1, CE_2,$ $CE_3$ ………

$$P1 = 0 \times .5 + 1 \times .5 = .5$$

$$P_2 = .5 \times .5 + 1 \times .5 = .75$$

$$P_3 = .75 \times .5 + 1 \times .5 = .8 \quad \text{etc.}$$

Step 6. Draw the graph of P vs CE and look up the P values corresponding to the relevant consequences of the Decision Tree. Let us say, we get the preference values as .03, .61, .63, .99 corresponding to the consequences of Rs. 40,000, Rs. 1,45,000, Rs. 1,50,000 and Rs. 4,95,000 respectively.

Step 7 We calculate the expected Preferences.

Expected Preference for Drilling, given that the test says oil

$=.818 \times .99 + 182 \times 0 = .809$

This is greater than the preference of not drilling, given that test says oil.

If test says oil, it is better to drill and expected preference in that case is .809.

Similarly, if test says no oil, expected preference of drilling (.174) is less than not drilling (.61). Hence if test says no oil, it is better not to drill and expected preference then is .61.

Expected Preference of taking test = .66 × .809 + .34 × .61 = .741. The Expected preference of not taking the test is given by :

.6 × 1 + .03 × .4 = .612.

Hence decision to take test will maximise his expected preference, i.e., in this case the decision is same as EMV maximising action. Though this need not always be true.

**Activity E**

Draw the Preference Curve for a decision maker who believes in maximising EMV. Consider another decision maker who is risk averse. Will the Preference Curve of the latter always be below that of the former? Justify your answer.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 8.6 OTHER APPROACHES

In the foregoing sections, we have assumed that the probabilities associated with the outcomes are known. In practice, we find situations where it is not possible to make any probability assessment. The EMV and preference criteria fail in such cases. The objective of this concluding section is to discuss some criteria that can be used under such circumstances.

Criteria when probability are not known.

a) **Criterion of Pessimism :**At the name suggests, the decision-making is based on pessimism, viz, the assumption that whatever alternative is chosen, the worst payoff corresponding to each alternative is actually going to occur. A rational criterion for decision-making in such a case is to maximise the minimum payoff.

b) **Criterion of Optimism :** A variant of (a), here, over and above the maximum of the nninumum payoff (say, $M_I$)), the maximum of the maximum payoff (say, $M_2$) is determined. Choosing would mean complete optimism (the opposite of choosing $M_M$). It is suggested that the d.m. find they maximum and minimum payoff for each alternative and then weigh them by his coefficient of optimism to arrive at the expected payoff for each alternative. The alternative with maximum expected payoff can then be chosen. Coefficient of Optimism lies between 0 and 1. It gives us the degree by which the maximum payoff is favoured by the d.m. vis-a-vis the minimum payoff.

c) **Criterion of Regret :** The criteria stems from the fact that a regret inbuilt-in in the decision-making, as the final decision on an alternative and the actual outcome after the decision has been taken, may not match, A regret of zero occurs when it matches. The regret can be measured as follows Consider our d.m. having two alternative investment proposals, the outcome corresponding to each proposal will be a failure or

Success depending on whether there is an economic depression or not. The consequences are as follows :

| Outcome<br>Alt | Depression | No Depression |
|---|---|---|
| 1 | –10 | 40 |
| 2 | – 6 | 20 |

Thus, if alternative 1 is chosen, and a depression actually occurs, then there is a cause for regret, as choosing 2 would have meant a loss of only 6 (vis-a-vis 10), thus regret = 10 - 6 - 4. Similarly, if there is no depression actually, and alt. 2 has been chosen, then a regret of 40-20 = occurs. Choosing alternative1 and later finding no depression would mean zero regret. Thus, the regret matrix is found:

| | D | ND |
|---|---|---|
| 1 | 4 | 0 |
| 2 | 0 | 20 |

Now, a pessimistic stand is taken and the criterion of minimising maximum regret is used for decision. For each alternative, the maximum regret is found, and finally the alternative with minimum value of maximum regret is chosen. Thus our d.m. would have chosen alternative 1.

d) **Subjectivists' Criterion :** The outcomes are assumed to be equally probable in this case, and EMV is used for decision. This is known as the subjectivists' stand.

The above four criteria are the best-known ones. Selection of the final criterion is purely subjective, as the obvious by now. However, each provides us with certain rationale and the d.m. can choose any,. depending on his own inclination.

ignou
THE PEOPLE'S
UNIVERSITY

**Activity F**

Consider the following problem where the decision maker has three alternative courses of action. Corresponding to each action there are possible outcomes, the probabilities of occurrence of which are unknown. The monetary payoff in each case is given in the matrix below :

| Outcomes / Actions | $0_1$ | $0_2$ | $0_3$ | $0_4$ |
|---|---|---|---|---|
| $A_1$ | 10 | 15 | 25 | 20 |
| $A_2$ | 30 | 20 | 45 | 15 |
| $A_3$ | 25 | 40 | 55 | 10 |

For example, if the decision maker chooses $A_1$, and the outcome $0_1$ occurs, he will get Rs. 10.

What will be the decision if the decision maker follows the criterion of pessimism? Will this decision change if he adopts the criterion of minimising the regret?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

ignou THE PEOPLE'S UNIVERSITY

## 8.7 SUMMARY

Decision Theory provides us with the framework and methods for analysing decision problems under uncertainty. A decision problem under uncertainty is characterised by different alternative courses of action and uncertain outcomes corresponding to each action. The problems can involve a single stage or a multi-stage decision process. Marginal Analysis is helpful in solving single stage problems, whereas the Decision Tree Approach is useful for solving multi-stage problems. In this unit we have examined how these methods can be applied to solve decision problems. While using these methods, we have used the criterion of maximising the Expected Monetary Value (EMV). Thus, EMV basically assumes that the decision maker is risk neutral. Preference Theory helps in incorporating the preference of the decision maker in the Decision Tree framework. We have seen how instead of maximising the EMV, we can maximise the expected preference, and thereby consider the decision maker's attitude towards risk. In the final section of this unit we have examined certain other criteria that are helpful in taking decisions, when the probabilities of occurrence of the outcomes are not known.

## 8.8 FURTHER READINGS

Raiffa, H., *Decision Analysis*, Addison-Wesley.

Schlairfer; R., *Analysis of Decisions under Uncertainty,* McGraw-Hill.

Schlairfer, R., *Probability and Statistics for Business Decision*, McGraw-Hill (Ch. 38)

Berry, W.L. et al., *Management Decision Sciences*, R.D. Irwin, Inc.: Homewood. (Ch. 5)

Miller, D.W. and M.K. Starr, *Executive Decisions and Operations Research*, Prentice-Hall: Englewood-Cliffs. (Chs. 1, 4, 5 & 6).

Giovanni Parmigiani and Curdes Inoue Decision Theory: Principles and Approaches, Wiley.

ignou
THE PEOPLE'S
UNIVERSITY

# BLOCK 3

# SAMPLING AND SAMPLING DISTRIBUTIONS

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 9 SAMPLING METHODS

### Objectives

On successful completion of this unit, you should be able to:

- appreciate why sampling is so common in managerial situations
- identify the potential sampling errors
- list the various sampling methods with their strengths and weaknesses
- distinguish between probability and non-probability sampling
- know when to use the proportional or the disproportional stratified sampling
- understand the role of multi-stage and multi-phase sampling in large sampling studies
- appreciate why and how non-probability sampling is used in spite of its theoretical weaknesses
- recognise the factors which affect the sample size decision.

### Structure

9.1 Introduction
9.2 Why Sampling?
9.3 Types of Sampling
9.4 Probability Sampling Methods
9.5 Non-Probability Sampling Methods
9.6 The Sample Size
9.7 Summary
9.8 Self-assessment Exercises
9.9 Further Readings

## 9.1 INTRODUCTION

Let us take a look at the following five situations to find out the common features among them, if any:

1) An inspector from the Weights &Measures department of the government goes to a unit manufacturing vanaspati. He picks up a small number of packed containers from the day's production, pours out the contents from each of these selected containers and weighs them individually to determine if the manufacturing unit is packing enough vanaspati in its containers to conform to what is claimed as the net weight in the label.

2) The personnel department of a large bank wants to measure the level of employee motivation and morale so that it can initiate appropriate measures to help improve the same. It administers a questionnaire to about 250 employees from different branches and offices all over India selected from a total of about 30,000 employees and analysis the

ignou
THE PEOPLE'S
UNIVERSITY

information contained in these 250 filled-in questionnaires to assess the morale and motivation levels of all employees.

3) The product development department of a consumer products company has developed a "new improved" version of its talcum powder. Before launching the new product, the marketing department gives a container of the old version first and after a week, a container of the new version to a group of 400 consumers and gets the feedback of these consumers on various attributes of the products. These consumer responses will form the basis for assessing the consumer perception of the new talcum powder as compared to the old talcum powder.

4) The quality control department of a company manufacturing fluorescent tubes checks the life of its products by picking up 15 of its tubes at random and letting them burn till each one of them fuses. The life of all its products is assessed based on the performance of these 15 tubes.

5) An industrial engineer takes 100 rounds of the shop floor over a period of six clays and based on these 100 observations, assesses the machine utilisation on the shop floor.

**What is Sampling**

On the face of it, there is little that is common among the five situations described above. Each one refers to a different functional area and the nature of the problem also is quite different from one situation to another. However, on closer observation, it appears that in all these situations one is interested in measuring some attribute of a large or infinite group of elements by studying only a part of that group. This process of inferring something about a large group of elements by studying only a part of it, is referred to as sampling.

Most of us use sampling in our daily life, e.g. when we go to buy provisions from a grocery. We might sample a few grains of rice or wheat to infer the quality of a whole bag of it. In this unit we shall study why sampling works and the various methods of sampling available so that we can make the process of sampling more efficient.

**Some Basic Concepts**

We shall refer to the collection of all elements about which some inference is to be made as the population. For example, in situation (ii) above,, the population is the set of 30,000 employees working in the bank and in situation (iii), the population comprises of all the consumers of talcum powder in the country.

We are basically interested in measuring some characteristics of the population. This could be the average life of a fluorescent tube, the percentage of consumers of talcum powder who prefer the "new improved" talcum powder to the old one or the percentage of time a machine is being used as in situation (v) above. Any characteristic of a population will be referred to as a **parameter** of the population.

In sampling, some population parameter is inferred by studying only a part of the population. We shall refer to the part of the population that has been chosen as a sample. Sampling, therefore, refers to the process of choosing a sample from the population so that some inference about the population can be made by studying the sample. For example, the sample in situation (ii) consists of the 250 employees from different branches and offices of the bank.

Any characteristic of a sample is called a **statistic**. For example, the mean life of the sample of 15 tubes in situation (iv) above is a sample statistic.

Conventionally, population parameters are denoted by Greek or capital letters and sample statistics by lower case Roman letters. There can be exceptions to this form of notation, e.g. population proportions is usually denoted by p and the sample proportion by p.

Figure I shows the concept of a population and a sample in the form of the Venn diagram, where the population is shown as the universal set and a sample is shown as a true subset of the population. The characteristics of a population and a sample and some symbols for these are presented in Table 1.

**Figure I: Population and Sample**

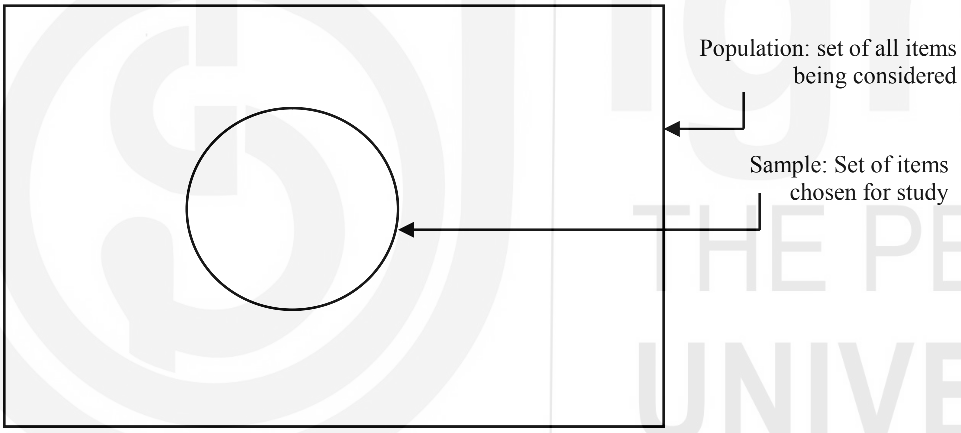

Population: set of all items being considered
Sample: Set of items chosen for study

**Table 1: Symbols for Population and Samples.**

| | POPULATION | SAMPLE |
|---|---|---|
| Characteristic | Parameter | Statistic |
| Symbols | Population size = N | Sample size = n |
| | Population mean = $\mu$ | Sample mean = $\overline{x}$ |
| | Population s.d. = $\sigma$ | Sample s.d. = s |
| | Population proportion = P | Sample proportion = p |

Sampling is not the only process available for making inferences about a population. For small populations, it may be feasible and practical, and sometimes desirable to examine every member of the population e.g. for inspection of some aircraft, components. This process is referred to as **census** or complete enumeration of the population.

## 9.2 WHY SAMPLING?

In the example situations given in section 9.1 above, the reasons for resorting to sampling should be very clear. We give below the various reasons which make sampling a desirable, and in many cases, the only course open for making an inference about a population.

### Time taken for the Study

Inferring from a sample can be much faster than from a complete enumeration of the population because fewer elements are being studied. In situation (iii) above in section 9.1, a complete enumeration of all consumers, even if feasible, would perhaps take so much time that it is unacceptable for product launch decisions.

### Cost involved for the Study

Sampling also helps in substantial cost reductions as compared to censuses and as we shall see later in this unit, a better sample design could reduce the cost of the study further. In many cases, like in situation (ii) above in section 9.1, it may be too costly, although feasible, to contact all the employees in the bank and get information from them.

### Physical Impossibility of Complete Enumeration

In many situations the element being studied gets destroyed while being tested. The fluorescent tubes in situation (iv) of section 9.1, which are chosen for testing their lives, get destroyed while being tested. In such cases, a complete enumeration is impossible as there would be no population left after such an enumeration.

### Practical Infeasibility of Complete Enumeration

Quite often it is practically infeasible to do a complete enumeration due to many practical difficulties. For example, in situation (iii) of section 9.1, it would be infeasible to collect information from all the consumers of talcum powder in India. Some consumers would have moved from one place to another during the period of study, some others would have stopped consuming talcum powder just before the period of study whereas some others would have been users of talcum powder during the period of study but would have stopped using it some time later. In such situations, although it is theoretically possible to do a complete enumeration, it is practically infeasible to do so.

### Enough Reliability of Inference based on Sampling

In many eases, sampling provides adequate information so that not much additional reliability can be gained with complete enumeration in spite of spending large amounts of additional money and time. It is also possible to quantify the magnitude of possible error on using; some types of sampling as will be explained later.

**Quality of Data Collected**

For large populations, complete enumeration also suffers from the possibility of spurious or unreliable data collected by the enumerators. On the other hand, there is greater confidence on the purity of the data collected in sampling as there can be better interviewing, better training and supervision of enumerators, better analysis of missing data and so on.

**Activity A**

When would you prefer complete enumeration to sampling?

…………………………………………………………………………………....

…………………………………………………………………………………....

…………………………………………………………………………………....

…………………………………………………………………………………....

…………………………………………………………………………………....

…………………………………………………………………………………....

**Activity B**

Name two decisions in each of the following functional areas, where sampling can be of use:

| **Functional Area** | **Decision** |
|---|---|
| Manufacturing | 1) Inspection of components<br>2) |
| Personnel | 1)<br>2) |
| Marketing | 1)<br>2) |
| Finance | 1)<br>2) |

## 9.3 TYPES OF SAMPLING

There are two basic types of sampling depending on who or what is allowed to govern the selection of the sample. We shall call them by the names of probability sampling and non-probability sampling.

**Probability Sampling**

In probability sampling the decision whether a particular element is included in the sample or not, is governed by chance alone. All probability sampling designs ensure that each element in the population has some nonzero probability of getting included in the sample. This would mean defining a procedure for picking up the sample, based on chance, and avoiding changes in the sample except by way of a pre-defined process again. The picking up of the sample is therefore totally insulated against the judgment, convenience or whims of any person involved with the study. That is why probability

ignou
THE PEOPLE'S
UNIVERSITY

sampling procedures tend to become rigorous and at times quite time¬consuming to ensure that each element has a nonzero probability of getting included in the sample. On-the other hand, when probability sampling designs are used, it is possible to quantify the magnitude of the likely error in inference made and this is of great help in many situations in building up confidence in the inference.

### Non-probability Sampling

Any sampling process which does not ensure some nonzero probability for each element in the population to be included in the sample would belong to the category of non-probability sampling. In this case, samples may be picked up based on the judgment or convenience of the enumerator. Usually, the complete sample is not decided at the beginning of the study but it evolves as the study progresses.

However, the very same factors which govern the selection of a sample e.g. judgment or convenience, can also introduce biases in the study. Moreover, there is no way that the magnitude of errors can be quantified when non-probability sampling designs are used.

Many times samples are selected by interviewers or enumerators "at random" meaning that the actual sample selection is left to the discretion of the enumerators. Such a sampling design would also belong to the non-probability sampling category and not the category of probability or random sampling.

## 9.4 PROBABILITY SAMPLING METHODS

In the category of probability sampling, we shall discuss the following four designs:

1) Simple Random Sampling
2) Systematic Sampling
3) Stratified Sampling
4) Cluster Sampling

One can also use sampling designs which are combinations of the above listed ones.

### Simple Random Sampling

Conceptually, simple random sampling is one of the simplest sampling designs and can work well for relatively small populations. However, there are many practical problems when one tries to use simple random sampling for large populations.

**What is simple random sampling?:** Suppose we have a population having N elements and that we want to pick up a sample of size n (< N). Obviously, there are many possible samples of size n.

Simple random sampling is a process which ensures that each of the samples of size n has an equal probability of being picked up as the chosen sample.

As we shall see later in this section this also implies that under simple random sampling, each element of the population has an equal probability of getting included in the sample.

All other forms of probability sampling use this basic concept of simple random sampling but applied to a part of the population at a time and not to the whole population.

Let us consider a small example to illustrate what simple random sampling is. Our population is a family of five members, two adults and three children, viz. A, B, C, D and E respectively. There are 10 different samples possible of size three as listed in Table 2 below. As we have shown in the same Table, if each of the 10 samples has an equal probability of 1/10 of being picked up, this implies that the probability that any particular element, say A or B, is included in the sample is the same.

In general, there are $\binom{N}{n}$ different samples of size n that can be picked up from a population of size N. Simple random sampling ensures that any of these samples has the same probability of picked up viz. $\frac{1}{\binom{N}{n}}$

**Table 2: Simple Random Sampling**

Population of size 5: (A, B, C, D and E)

Let P [ABC] be the probability that the sample of size 3 containing elements A, B and C, is chosen.

Simple Random Sampling ensures that

P[ABC] =1/10
P[ABD] =1/10
P[ABE] =1/10
P[ACD] =1/10
P[ACE] =1/10

P[ADE] =1/10
P[BCD] =1/10
P[BCE] =1/10
P[BDE] =1/10
P[CDE] =1/10

∴ Probability that element A

is in the sample, P(A) = P[ACC] + P[ABD] + P[ABE] + P[ACD] + P[ACE] + [ADE] = 6/10

and P(B) = P[ABC] + P[ABD] + P[ABE] + P[BCD] + P[BCE] + P[BDE]= 6/10

Similarly P(C) = 6/10

P(D) = 6/10

and P(E) = 6/10

If we want to find the probability that element A (or any other element for that matter) is included in the sample picked up, we have to find the number of different samples in which this element A occurs. There are (n -1) positions available in the sample (since one is occupied by A) which can be picked up from any of the (N-1) elements of the population (since A is not available to be picked up) and so there are:

$\binom{N}{n}$ different samples in which element A occurs

Therefore, the probability that element A is included in

$$\text{the sample} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!}{(n-1)!(N-n)!} \times \frac{n!(N-n)!}{N!}$$

$$= \frac{n}{N}$$

The fact that every element of the population has an equal probability of getting included in the sample is made use of in actually picking up simple random samples.

**Sampling with and without replacement**: We have implicitly assumed above that we are sampling without replacement, i.e. if an element is picked up-once, it is not available to be picked up again. This is how most practical samples are, but as a concept, it is possible to think in terms of sampling with replacement in which case an element, after being picked up and included in the sample, is replaced in the population so that it can be picked up again.

What is important for us to note at this stage is that even in the case of simple random sampling with replacement, each element has an equal probability of getting included in the sample.

**How is simple random sampling done?:** It is imperative to have a list of all the members of the population before a simple random sample can be picked up. Such an exhaustive list of all population members is called a sampling frame.

Suppose we write the name of one such member on a chit of paper and thus have N chits in a bowl, one chit for each member of the population. We can then mix the chits well and pick up one chit at random to represent one member of the sample. If we want a sample of size n, we have to repeat this process n times and we shall have a simple random sample of size n consisting of the names of members appearing on the chits picked.

It is easy to see that if we replace the chits in the bowl after noting down the name of the element, we will have a simple random sample with replacement and one without replacement if we do not.

As the population size increases, it becomes more and more difficult to work with chits and one can simulate this process on a computer or by using a table of random numbers. We can associate a serial number with each member of our population and then instruct a computer to pick up a member from 1 through N using its pseudo¬random number generator. This ensures that every number from 1 through N has an equal probability of getting picked up and so the sample selected is a simple random sample.

ignou THE PEOPLE'S UNIVERSITY

We can also use a table of random numbers to pick up a simple random sample. In a table-of random numbers there is an equal probability for any digit from 0 to 9 to appear in any particular position. In table 3 we have a page of five digit random numbers containing 100 such numbers. The most important thing in using a random number table is to specify to the minutest detail the sequence of steps that has been decided before the table is actually referred to. We shall demonstrate this with an example.

Suppose we have a population of size 900 with each number being given a serial number ranging from 000 through 899 and we want to pick up a simple random sample of size 20. We proceed by defining a procedure.

1) Starting point and direction of movement. We may decide to start with the top left hand number and consider the first three digits (from left) as the three-digited random number picked up e.g. the first number would then be 121. We also specify that we shall move down a column to pick up further numbers-e.g. the second number would be 073, If there is no further number down the column, we shall go to the top of the next column of five-digited numbers and pick up the first three digits (from left)-e.g. after 851 our next number shall be 651.
2) Checking the number picked up. If the number picked up is in the range 000 to 899, we accept the number but if it is outside this range, we shall discard it and pick up the next number-e.g. after the third number 703, we discard 934 and the fourth member of the sample would be 740. Similarly, if we are doing sampling without replacement and a number is picked up again, it is discarded and we move to the next three-digited number.

   Using this process, if we want a sample of size 10, our sample would contain members with the following numbers: 121, 073, 703, 740, 736, 513, 464, 571, 379 and 412.

**Simple random sampling in practice**: Simple random sampling, as described here, is not the most efficient sampling design either statistically or economically in all practical situations. However, it forms the basis for all other forms of probability sampling which are used on parts of the population or sub-population and not on the population as a whole.

**Table 3: Table of five-digited random numbers**

| | | | | |
|---|---|---|---|---|
| 12135 | 65186 | 86886 | 72976 | 79885 |
| 07369 | 49031 | 45451 | 10724 | 95051 |
| 70387 | 53186 | 97116 | 32093 | 95612 |
| 93451 | 53493 | 56442 | 67121 | 70257 |
| 74077 | 66687 | 45394 | 33414 | 15685 |
| 73627 | 54287 | 42596 | 05544 | 76826 |
| 51353 | 56404 | 74106 | 66185 | 23145 |
| 46426 | 12855 | 48497 | 05532 | 36299 |
| 57126 | 99010 | 29015 | 65778 | 93911 |

ignou THE PEOPLE'S UNIVERSITY

| | | | | |
|---|---|---|---|---|
| 37997 | 89034 | 79788 | 94676 | 32307 |
| 41283 | 42498 | 73173 | 21938 | 22024 |
| 76374 | 68251 | 71593 | 93397 | 26245 |
| 51668 | 47244 | 13732 | 48369 | 60907 |
| 17698 | 32685 | 24490 | 56983 | 81152 |
| 12448 | 00902 | 07263 | 16764 | 71261 |
| 52515 | 93269 | 61210 | 55526 | 71912 |
| 43501 | 10248 | 34219 | 83416 | 91239 |
| 45279 | 19382 | 82151 | 57365 | 84915 |
| 11437 | 98102 | 58168 | 61534 | 69495 |
| 85183 | 38161 | 22848 | 06673 | 35293 |

As mentioned earlier, in listing all members of the population viz. a frame is required before a simple random sample can be chosen. In many situations the frame is not available nor is it practical to prepare the frame in a time and cost-effective manner. Obviously, under such conditions simple random sampling is not a viable sampling design.

Most large populations are not homogeneous and can be broken down into more homogeneous units. In such conditions one can design sampling schemes which are statistically more efficient, meaning that they allow the same precision from smaller sample sizes.

Similarly by picking up members from geographically closer areas the cost efficiency of the sampling design can be improved. Cluster sampling is based on this concept.

The process of picking up a simple random sample through using a table of random numbers or any other such aids as discussed earlier, is rather cumbersome and not very purposeful to the uninitiated interviewer. Simpler forms of sampling overcomes this handicap of simple random sampling.

ignou THE PEOPLE'S UNIVERSITY

**Activity C**

There are 20 elements in a population, each identified by a letter of the English alphabet from A through T. Using the random number table given in. Table 3, describe how you would pick up a sample of size 5 when sampling is done without replacement.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## Systematic Sampling

Systematic sampling proceeds by picking up one element after a fixed interval depending on the sampling ratio. For example, if we want to have a sample of size 10 from a population of size 100, our sampling ratio would be n/N = 10/100 = 1/10. We would, therefore, have to decide where to start from among the first 10 names in our frame. If this number happens to be 7 for example, then the sample would contain members having serial numbers 7,17,27,….97 in the frame. It is to be noted that the random process establishes only the first member of the sample the rest are pre-ordained automatic because of the known sampling ratio.

Systematic sampling in the previous example would choose one out of ten possible samples each starting with either number 1, or number 2, or ....number 10. This is usually decided by allowing chance to play its role e.g. by using a table of random numbers.

Systematic sampling is relatively much easier to implement compared to simple random sampling. However, there is one possibility that should be guarded against while using systematic sampling-the possibility of a strong bias in the results if there is any periodicity in the frame that parallels the sampling ratio. One can give some ridiculously simple example to highlight the point. If you were making studies on the demand for various banking transactions in a bank branch by studying the demand on some days randomly selected by systematic sampling, be sure that your sampling ratio is not 1/7 or 1/14 etc. Otherwise you would always be studying the demand on the same day of the week and your inferences could be biased depending on whether the day selected is a Monday or a Friday and so on. Similarly, when the frame contains addresses of flats in buildings all alike and having say 12 flats in one building, systematic sampling with a sampling ratio of 1/6, 1/60 or any other such fraction would bias your sample with flats of only one type e.g. a ground floor corner flat i.e., all types of flats would not be members of your sample; and this might lead to biases in the inference made.

If the frame is arranged in an order, ascending or descending, of some attribute then the location of the first sample element may affect the result of the study. For example, if our frame contains a list of students arranged in a descending order of their percentage in the previous examination and we are picking a systematic sample with a sampling ratio of 1/50. If the first number picked is 1 or 2, then the sample chosen will be academically much better off compared to another systematic sample with the first number chosen as 49 or 50. In such situations, one should devise ways of nullifying the effect of bias due to starting number by insisting on multiple starts after a small cycle or other such means.

On the other hand, if the frame is so arranged that similar elements are grouped together, then systematic sampling produces almost a proportional stratified sample and would be, therefore, more statistically efficient than simple random sampling.

Systematic sampling is perhaps the most commonly used method among the probability sampling designs and for many purposes e.g. for estimating the precision of the results, systematic samples are treated as simple random samples.

### Stratified Sampling

Stratified sampling is more complex than simple random sampling, but where applied properly, stratification can significantly increase the statistical efficiency of sampling.

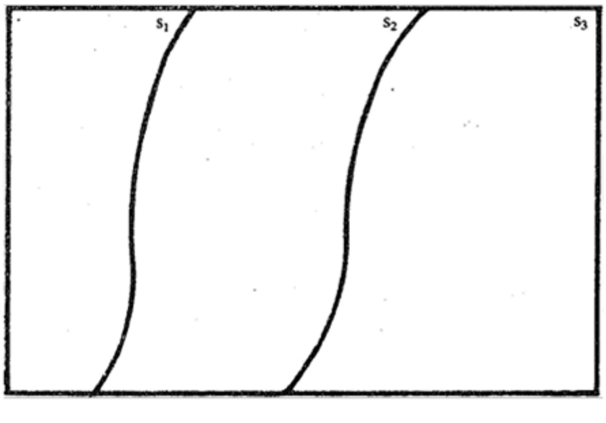
**The concept:** Suppose we are interested in estimating the demand of non-aerated beverages in a residential colony. We know that the consumption of these beverages has some relationship with the family income and that the families residing in this colony can be classified into three categories-viz., high income, middle income and low income families. If we are doing a sampling study we would like to make sure that our sample does have some members from each of the three categories-perhaps in the same proportion as the total number of families belonging to that category-in which case we would have used proportional stratified sampling. On the other hand, if we know that the variation in the consumption of these beverages from one family to another is relatively large for the low income category whereas there is not much variation in the high income category, we would perhaps pick up a smaller than proportional sample from the high income category and a larger than proportional sample from-the low income category. This is what is done in disproportional stratified sampling.

The basis for using stratified sampling is the existence of strata such that each stratum is more homogeneous within and markedly different from another stratum. The higher the homogeneity within each stratum, the higher the gain in statistical efficiency due to stratification.

**What are strata?:** The strata are so defined that they constitute a partition of the population-i.e., they are mutually exclusive and collectively exhaustive. Every element of the population belongs to one stratum and not more than one stratum, by definition. This is shown in Figure II in the form of a Venn diagram, where three strata have been shown.

A stratum can therefore he conceived of as a sub-population which is more homogeneous than the complete population-the members of a stratum, are similar to each other and are different from the members of another stratum in the characteristics that we are measuring.

**Figure II: A Population with three strata**



ignou
THE PEOPLE'S
UNIVERSITY

**Proportional stratified sampling:** After defining the strata, a simple random sample is picked up from each of the strata. If we want to have a total sample of size 100, this number is allocated to the different strata-either in proportion to the size of the stratum in the population or otherwise.

If the different strata have similar variances of the characteristic being measured, then the statistical efficiency will be the highest if the sample sizes for different strata are in the same proportion as the size of the respective stratum in the population. Such a design is called proportional stratified sampling and is shown in Table 4 below.

If we want to pick up a proportional stratified sample of size n from a population of size N, which has been stratified to p different strata with sizes $N_{1,}N_2, \ldots \ldots \ldots \ldots . N_p$ respectively, then the sample sizes for different strata, viz $n_1, n_2, \ldots \ldots n_p$ will be given by

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_p}{N_p} = \frac{n}{N}$$

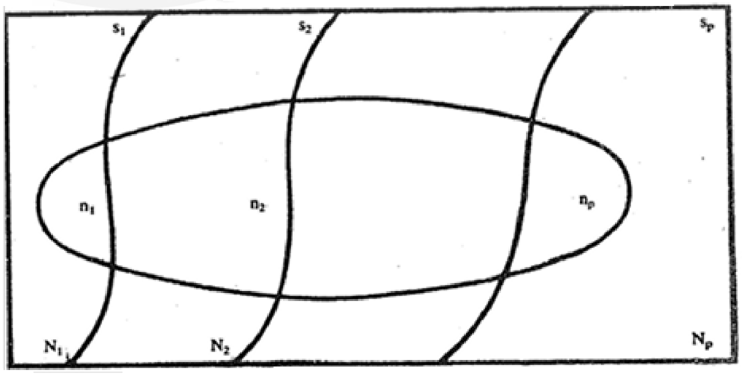
**Table 4: Proportional Stratified Sampling**

| Stratum No. (i) | No. of Elements in stratum ($N_i$) | Sample size ($n_i$) | Sampling Ratio ($n_i/N_i$) |
|---|---|---|---|
| 1 | 200 | 10 | 1/20 |
| 2 | 300 | 15 | 1/20 |
| 3 | 500 | 25 | 1/20 |
| Total | 1000 | 50 | 1/20 |

ignou THE PEOPLE'S UNIVERSITY

The strata and the samples from each stratum are shown in the form of a Venn diagram in Figure III below, where $S_1$, $S_2$ etc. refer to the stratum number 1 and stratum number 2 etc. respectively.

**Figure III: Stratified Sampling**



**Disproportional stratified sampling**: If the different strata in the population have unequal variances of the characteristic being measured, then the sample size allocation decision should consider the variance as well. It would be logical to have a smaller sample from a stratum where the variance is smaller than from another stratum where the variance is higher. In fact, if

$\sigma_1^2, \sigma_2^2, \dots\dots, \sigma_p^2$ are the variance of the p strata respectively, then the statistical efficiency is the highest when

$$\frac{n_1}{N_1\sigma_1} = \frac{n_2}{N_2\sigma_2} = \cdots = \frac{n_p}{N_r\sigma_p}$$

where the other symbols have the same meaning as in the previous example.

Suppose the variances of the characteristic we are measuring were different for each of the three strata of the earlier example and were actually as shown in Table 5. If the total sample size was still restricted to 50, the statistically optimal allocation would be as given in Table 5 and one can compare this Table with Table 4 above to find that the sampling ratio would fall for Stratum-3 as the variance is smaller here and would go up for Stratum-2 where the variance is larger.

**Table 5: Disproportional Stratified Sampling**

| Stratum No.(i) | No. of Elements in Stratum ($\mathbf{N}_i$) | Stratum Variance ($\sigma_i^2$) | Stratum s.d. ($\sigma_i$) | Sample size ($n_i$) | Sampling Ratio ($n_i/N_i$) |
|---|---|---|---|---|---|
| 1 | 200 | 2.25 | 1.5 | 13 | 0.065 |
| 2 | 300 | 4.00 | 2.0 | 26 | 0.087 |
| 3 | 500 | 0.25 | 0.5 | 11 | 0.022 |
| Total | 1000 | | | 50 | |

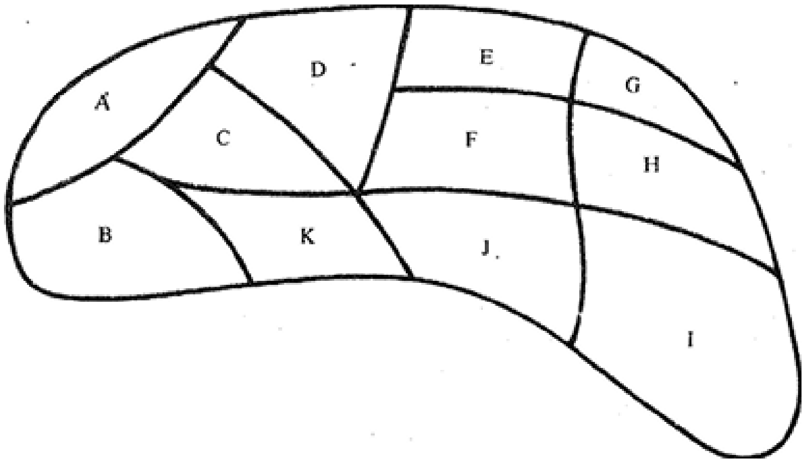
**Stratified sampling in practice**: Stratification of the population is quite common in managerial applications because it also allows to draw separate conclusions for each stratum. For example, if we are estimating the demand for a non-aerated beverage in a residential colony and have stratified the population based on the family income, then we would have data pertaining to each stratum which might be useful in making many marketing decisions.

Stratification requires us to identify the strata such that the intra-stratum differences are as small as possible and inter-strata differences as large as possible. However, whether a stratum is homogeneous or not-in the characteristic that we are measuring e.g. consumption of non-aerated beverage in the family in the previous example-can be known only at the end of the study whereas stratification is to be done at the beginning of the study and that is why some other variable like family income is to be used for stratification. This is based on the implicit assumption that family income and consumption of non-aerated beverages are very closely associated with each other. If this assumption is true, stratification would increase the statistical efficiency of sampling. In many studies, it is not easy to find such associated variables which can be used as the basis for stratification and then stratification may not help in increasing the statistical efficiency, although the cost of the study goes up due to the additional costs of stratification.

### Cluster Sampling

Let us take up the situation where we are interested in estimating the demand for a non-aerated beverage in a residential colony again. The colony is divided into 11 blocks, called Block A through Block K as shown in Figure IV below.

**Figure IV: Blocks in a residential colony**



We might use cluster sampling in this situation by treating each block as a cluster. We will then select 2 blocks out of the 11 blocks at random and then collect information from all families residing in those 2 blocks.

**Cluster vs stratum:** We can now compare cluster sampling with stratified sampling. Stratification is done to make the strata homogeneous within and different from other strata. Clusters, on the other hand, should be heterogeneous within and the different clusters should be similar to each other. A clusture, ideally, is a mini-population and has all the features of the population.

The criterion used for stratification is a variable which is closely associated with the characteristic we are measuring e.g. income level when we are measuring the family consumption of non-aerated beverages in the example quoted earlier. On the other hand, convenience of data collection is usually the basis for cluster definitions.

Geographic contiguity is quite often used for clusture definitions, like in Figure IV above and in such cases, cluster sampling is also known as Area Sampling.

There are very fewer strata and one requires to pick up a random sample from each of the strata for drawing inferences. In cluster sampling, there are many clusters out of which only a few are picked up by random sampling and then the clusters are completely enumerated.

**Cluster sampling in practice:** Cluster sampling is used primarily because it allows for great economies in data collection costs since the travel related costs etc. are smaller. Although it is statistically less efficient than simple random sampling in most cases, this deficiency may be more than offset by the high economic efficiency that it offers. For example, to get a certain precision level one might need a sample size of 100 under simple random sampling and a sample size of 175 under cluster sampling. However if the cost of data collection is Rs. 20 under simple random sampling and only Rs. 5 under cluster sampling, it would be cost-effective to use cluster sampling.

ignou THE PEOPLE'S UNIVERSITY

Cluster sampling is rarely used in single-stage sampling plans. In a national survey, a district might be treated as a cluster and cluster sampling used in the first stage to pick up 15 districts in the country. Some other form of probability sampling like stratified sampling cluster sampling etc. is then used to go to a smaller sampling unit.

If a frame has to be developed, then cluster sampling allows us to save on the cost of developing a frame because frames need to be developed only for the selected clusters and not for the whole population.

### Multi-stage and Multi-phase Sampling

In most large surveys one uses multi-stage sampling where the sampling unit is something larger than an individual element of the population in all stages but the final. For example, in a national survey on the demand of fertilizers one might use stratified sampling in the first stage with a district as a sampling unit and the average rainfall in the district as the criterion for stratification. Having obtained 20 districts from this stage, cluster sampling may be used in the second stage to pick up 10 villages in each of the selected districts. Finally, in the third stage, stratified sampling may be used in each village to pick up frames in each of the strata defined with land holding as the criterion.

Multi-phase sampling, on the other hand, is designed to make use of the information collected in one phase to develop a sampling design in a subsequent phase. A study with two phases is often called **double sampling**. The first phase of the study might reveal a relationship between the family consumption of non-aerated beverages and the family income and this information would then be used in the second phase to stratify the population with family income as the criterion.

### Activity D

Using a calendar for the current year, identify a systematic sample of size 10 when the sampling ratio is 1/20. (Tomorrow is the first possible member of the sample.)

…………………………………………………………………………………....

…………………………………………………………………………………....
…………………………………………………………………………………....
…………………………………………………………………………………....

### Activity E

A lot of debate is going on regarding the grant of statehood to Delhi. If you plan to do a sample survey of 3000 residents in Delhi on this question, what kind of sampling design would you use? In Delhi, many colonies are posh and many others are poor and you believe that the response on statehood is highly dependent on the income level of the respondent.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 9.5 NON-PROBABILITY SAMPLING METHODS

Probability sampling has some theoretical advantages over non-probability sampling. The bias introduced due to sampling could be completely eliminated and it is possible to set a confidence interval for the population parameter that is being studied. In spite of these advantages of probability sampling, non-probability sampling is used quite frequently in many sampling surveys. This is so because all are based on practical considerations.

Probability sampling requires a list of all the sampling units and this frame is not available in many situations nor is it practically feasible to develop a frame of say all the households in a city or zone or ward of a city. Sometimes the objective of the study may not be to draw a statistical inference about the population but to get familiar with extreme cases or other such objectives. In a dealer survey, our objective may be to get familiar with the problems faced by our dealers so that we can take some corrective actions, wherever possible. Probability sampling is rigorous and this rigour e.g. in selecting samples, adds to the cost of the study. And finally, even when we are doing probability sampling, there are chances of deviations from the laid out process especially where some samples are selected by the interviewers at site-say after reaching a village. Also, some of the sample members may not agree to be interviewed or not available to be interviewed and our sample may turn out to be a non-probability sample in the strictest sense of the term.

### Convenience Sampling

In this type of non-probability sampling, the choice of the sample is left completely to the convenience of the interviewer. The cost involved in picking up the sample is minimum and the cost of data collection is also generally low, e.g. the interviewer can go to some retail shops and interview some shoppers while studying the demand for non-aerated beverages.

However, such samples can suffer from excessive bias from known or unknown sources and also there is no way that the possible errors can be quantified.

### Purposive Sampling

Inconvenience sampling, any member of the population can be included in the sample without any restriction. When some restrictions are put on the possible inclusion of a member in the sample, the sampling is called purposive.

**Judgment Sampling:** In judgment sampling, the judgment or opinion of some experts forms the basis for sample selection. The experts are persons who are believed to have information on the population which can help in giving us better samples. Such sampling is very useful when we want to study rare events, or when members have extreme positions, or even when the objective of the study is to collect a wide cross-section of views from one extreme to the other.

**Quota Sampling:** Even when we are using non-probability sampling, we might want our sample to be representative of the population in some defined ways. This is sought to be achieved in quota sampling so that the bias introduced by sampling could be reduced.

If in our population, 20% of the members belong to the high income group, 30% to the middle income group and 50% to the low income group and we are using quota sampling, we would specify that the sample should also contain members in the same proportion as in the population e.g. 20% of the sample members would belong to the high income group and so on.

The criteria used to set quotas could be many. For example, family size could be another criterion and we can set quotas for families with family size upto 3, between 4 and 5, and above 5. However, if the number of such criteria is large, it becomes difficult to locate sample members satisfying the combination of the criteria. In such cases, the overall relative frequency of each criterion in the sample is matched with the overall relative frequency of the criterion in the population.

## 9.6 THE SAMPLE SIZE

How large a sample should be taken in a study? So far in this unit we have not addressed ourselves to this question. At this stage, we will only mention some factors affect the sample size decision and in later units some of these ideas will be gone into in more depth.

One of the most important factors that affect the sample size is the extent of variability in the population. Taking an extreme case, if there is no variability, i.e. if all the members of the population are exactly identical, a sample of size 1 is as good as a sample of 100 or any other number. Therefore, the larger the variability, the larger is the sample size required.

A second consideration is the confidence in the inference made-the larger the sample size the higher is the confidence. In many situations, the confidence level is used as the basis to decide sample size as we shall see in the next unit.

In many real life situations, the factor of overriding importance is the cost of the study and the problem then becomes one of designing a sampling scheme to achieve the highest statistical efficiency subject to the budget for the study. It is here that cluster sampling and convenience sampling score over other more statistically efficient methods of sampling, since the unit cost of data collection is lower.

## 9.7 SUMMARY

In this unit we have looked at various sampling methods available when one wants to make some inferences about a population without enumerating it completely. We started by looking at some situations where sampling was being done and then found that in many situations sampling may be the only feasible way of knowing something about the population-either because of the time or cost involved, or because of the physical impossibility or practical infeasibility of observing the complete population. Also, sampling can give us adequate results in many applications and can be preferred over complete enumeration as it ensures a higher purity of the data collected, especially when the population is large.

We noted that there are two basic methods of sampling-probability sampling which ensures that every member of the population has a calculable nonzero probability getting included in the sample and non-probability sampling where there is no such assurance. Probability sampling is theoretically superior to non-probability sampling as it helps us in reducing the bias and also allows us to quantify the possible error involved, but non-probability sampling is less rigorous, easy to use, practically feasible and gives adequate results in some applications.

Among the probability sampling methods, simple random sampling works the best when the population is homogeneous but may have many practical limitations when the population is large. Simple random sampling ensures that each of the possible samples of a particular size has an equal probability of getting picked up as the sample selected and it also implies that each element of the population has an equal probability of being included in the sample. Systematic sampling starts with a random start and picking up members after a fixed interval down a list of all members called the sampling frame. If the population can be broken down into smaller, more homogeneous sub-populations or strata, then stratified sampling should be used which allows higher economic efficiency as the cost of data collection per element is reduced if members are physically or otherwise closer to each other as they are. in a cluster. Most large studies are based on multi-stage sampling where different sampling methods are used at each stage. In some studies multi-phase sample is also used, especially where the information collected in one phase is used in the sampling design of a later phase.

We have also discussed some of the non-probability sampling methods used in practice. If any member of the population could be included in the sample, we would get a convenience sample. On the other hand, if the entry is subject to the judgment of some expert or experts who have a better knowledge of the population, we would have used judgment sampling and if the sample is made representative of the population by setting quotas for elements satisfying different criteria, this is called quota sampling. Purposive sampling is a genuine name for all non-probability sampling methods where restrictions are used on entry. We have looked at all of these sampling methods to gauge their strengths and weaknesses and also to find their applicability under different conditions.

## 9.8 SELF-ASSESSMENT EXERCISES

1) List the various reasons that make sampling so attractive in drawing conclusions about the population.

2) What is the major difference between probability and non-probability sampling?

3) A study aimes to quantify the organisational climate in any organisation by administering a questionnaire to a sample of its employees. There are 1000 employees in a company with 100 executives, 200 supervisors and 700 workers. If the employees are **stratified** based on this classification and a sample of 100 employees is required, what should the sample size be from each **stratum**, if proportional **stratified** sampling is used?

4) In question 3 above, if it is known that the standard deviation of the response for Qfexecutives is 1.9, for supervisors is 3.2 and for workers is 2.1, what should the respective sample sizes be?

   Please state for each of the following statements, which of the given response is the most correct:

5) To determine the salary, the sex and the working hours structure in a large multi¬storeyed office building, a survey was conducted in which all the employees working on the third, the eighth and the thirteenth floors were contacted. The sampling scheme used was:
   i) simple random sampling
   ii) stratified sampling
   iii) cluster sampling
   iv) convenience sampling

6) We do not use extremely large sample sizes because
   i) the unit cost of data collection and data analysis increases as the sample size increases-e.g. it costs more to collect the thousandth sample member as compared to the first.
   ii) the sample becomes unrepresentative as the sample size is increased.
   iii) it becomes more difficult to store information about large sample size.
   iv) As the sample size increases, the gain in having an additional sample element falls and so after a point, is less than the cost involved in having an additional sample element:

7) If it is known that a population has groups which have a wide amount of variation within them, but only a small variation among the groups themselves, which of the following sampling schemes would you consider appropriate:
   i) cluster sampling
   ii) stratified sampling
   iii) simple random sampling
   iv) systematic sampling

8) One of the major drawbacks of judgement sampling is that

i) the method is cumbersome and difficult to use

ii) there is no way of quantifying the magnitude of the error involved

iii) it depends on only one individual for sample selection

iv) it gives us small sample sizes

## 9.9 FURTHER READINGS

Levin, R.I;, *Statistics for Management*, Prentice Hall of India: New Delhi..

Mason, R.D., *Statistical Techniques in Business and Economics*, Richard D. Irwin, Inc: Homewood.

Mendenhall, W.,R.L. Scheaffer and D.D. Wackerly, *Mathematical Statistics with Applications,* Danbury Press: Boston.

Plane, D.R. and E.B. Oppermann, *Business and Economic Statistics*; Business Publications, Inc: Plano.

Bruce Bowerman. Business statistics for practice, McGraw Hill.

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 10 SAMPLING DISTRIBUTIONS

**Objectives**

When you have successfully completed this unit, you should be able to:

- understand the meaning of sampling distribution of a sample statistic
- obtain the sampling distribution of the mean
- get an understanding of the sampling distribution of variance
- construct the sampling distribution of the proportion
- know the Central Limit Theorem and appreciate why it is used so extensively in practice
- develop confidence intervals for the population mean and the population proportion
- determine the sample size required while estimating the population mean or the population proportion.

**Structure**

10.1 Introduction
10.2 Sampling Distribution of the Mean
10.3 Central Limit Theorem
10.4 Sampling Distribution of the Variance
10.5 The Student's t Distribution
10.6 Sampling Distribution of the Proportion
10.7 Interval Estimation
10.8 The Sample Size
10.9 Summary
10.10 Self-assessment Exercises
10.11 Further Readings

## 10.1 INTRODUCTION

Having discussed the various methods available for picking up a sample from a population we would naturally be interested in drawing inferences about the population based on our observations made on the sample members. This could mean estimating the value of a population parameter, testing a statistical hypothesis about the population, comparing two or more populations, performing correlation and regression analysis on more than one variable measured on the sample members, and many other inferences. We shall discuss some of these problems in this and the subsequent units.

### What is a Sampling Distribution?

Suppose we are interested in drawing some inference regarding the weight of containers produced by an automatic filling machine. Our population, therefore, consists of all the filled-containers produced in the past as well as those which are going to be produced in the future by the automatic filling machine. We pick up a sample of size n and take measurements regarding the characteristic we are interested in viz. the weight of the filled container on each of our sample members. We thus end up with n sample values $x_1, x_2, \ldots \ldots \ldots x_n$. As described in the previous unit, any quantity which can be determined as a function of the sample values $x_1, x_2, \ldots, x_n$ is called a sample statistic.

Referring to our earlier discussion on the concept of a random variable, it is not difficult to see that any sample statistic is a random variable and, therefore, has a probability distribution or a probability density function. It is also known as the sampling distribution of the statistic. In practice, we refer to the sampling distributions of only the commonly used sampling statistics like the sample mean, sample variance, sample proportion, sample median etc., which have a role in making inferences about the population.

### Why Study Sampling Distributions?

Sample statistics form the basis of all inferences drawn about populations. If we know the probability distribution of the sample statistic, then we can calculate the probability that the sample statistic assumes a particular value (if it is a discrete random variable) or has a value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences. We will demonstrate this by an example.

Suppose we know that 45% of the population of all users of talcum powder prefer our brand to the next competing brand. A "new improved" version of our brand has been developed and given to a random sample of 100 talcum powder users for use. If 60 of these prefer our "new improved" version to the next competing brand, what should we conclude? For an answer, we would like to know the probability that the sample proportion in a sample of size 100 is as large as 60% or higher when the true population proportion is only 45%, i.e. assuming that the new version is no better than the old. If this probability is quite large, say 0.5, we might conclude that the high sample proportion viz. 60% is perhaps because of sampling errors.and the new version is not really superior to the old. On the other hand, if this probability works out to a very small figure, say 0.001, then rather than concluding that we have observed a rare event we might conclude that the true population proportion is higher than 45%, i.e. the new version is actually superior to the old one as perceived by members of the population. To calculate this probability, we need to know the probability distribution of sample proportion or the sampling distribution of the proportion.

## 10.2 SAMPLING DISTRIBUTION OF THE MEAN

We shall first discuss the sampling distribution of the mean. We start by discussing the concept of the sample mean and then study its expected value and variance in the general case. We shall end this section by describing the sampling distribution of the mean in the special case when the population distribution is normal.

### The Sample Mean

Suppose we have a simple random sample of size n picked up from a population. We take measurements on each sample member in the characteristic of our interest and denote the observation as $\mathrm{x_1, x_2, \dots, x_n}$ respectively. The sample mean for this sample, represented by x, is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

If we pick up another sample of size n from the same population, we might end tip a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance causes. The distribution of the sample mean is also referred to as the sampling distribution of the mean.

However, to observe the distribution of $\bar{x}$ empirically, we have to take many samples of size n and determine the value of $\bar{x}$ for each sample. Then, looking at the various observed values of z, it might be possible to get an idea of the nature of the distribution.

### Sampling from Infinite Populations

We shall study the distribution of z in two cases-one when the population is finite and we are sampling without replacement; and the other when the population is infinitely large or when the sampling is done with replacement. We start with the latter.

We assume we have a population which is infinitely large and having a population mean of and a population variance of $\sigma^2$. This implies that if x is a random variable denoting the measurement of the characteristic that we are interested in, on one element of the population picked up randomly, then

the expected value of x, E(x) = $\mu$

and the variance of x, Var (x) = $\sigma^2$

The sample mean, $\bar{X}$, can be looked at as the sum of n random variables, viz $\mathrm{x_1, x_2, \dots, x_n}$ each being divided by (1/n). Here $\mathrm{X_1}$ is a random variable representing the first observed value in the sample, $\mathrm{X_2}$ is a random variable representing the second observed value and so on. Now, when the population is infinitely large, whatever be the value of $\mathrm{X_1}$, the distribution of $\mathrm{X_2}$ is not affected by it. This is true of any other pair of random variables as well. In other words$x_1, x_2, \dots, x_n$ are independent random variables and all are picked up from the same population.

$$\therefore E(x_1) = \mu \text{ and Var } (x_1) = \sigma^2$$

$E(x_2) = \mu$ and Var $(x_2) = \sigma^2$ and so on

Finally,

$$E(\bar{x}) = E\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right]$$

$$= \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + \cdots + \frac{1}{n}E(x_n)$$

$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu$$

$$= \mu.$$

and Var $(\bar{x})$ = Var $\left[\frac{x_1+x_2+\cdots+x_n}{n}\right]$

$$= \text{Var}\left(\frac{x_1}{n}\right) + \text{Var}\left(\frac{x_2}{n}\right) + \cdots + \text{Var}\left(\frac{x_n}{n}\right)$$

$$= \frac{1}{n^2}\text{Var }(x_1) + \frac{1}{n^2}\text{Var }(x_2) + \cdots + \frac{1}{n^2}\text{Var }(x_n)$$

$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2$$

$$= \frac{\sigma^2}{n}$$

We have arrived at two very important results for the case when the population is infinitely large, which we shall be using very often. The first says that the expected value of the sample mean is the same as the population mean while the second says that the variance of the sample mean is the variance of the population divided by the sample size.

If we take a large number of samples of size n, then the average value of the sample means tends to be close to the true population mean. On the other hand, if the sample size is increased then the variance of gets reduced and by selecting an appropriately large value of n, the variance of x can be made as small as desired.

Thee standard deviation of x is also called the standard error of the mean. Very often we estimate the population mean by the sample mean. The standard error of the mean indicates the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we are reasonably confident that whatever sample mean value we have observed cannot be very far away from the true value.

The standard error of the mean is represented by $\sigma_x$.

### Sampling With Replacement

The above results have been obtained under the assumption that the random variables $X_1, X_2, \ldots, X_n$, are independent. This assumption is valid when the population is infinitely large. It is also valid when the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. Hence, if the sampling is done with replacement, we would again have

$E(\bar{x}) = \mu$

and Var $(\bar{x}) = \frac{\sigma^2}{n}$

i.e. $\sigma_{\bar{x}} = \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$

### Sampling Without Replacement from Finite Populations

When a sample is picked up without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on. As the n random variables representing the n sample members do not remain independent, the expression for the variance of x changes. We are only mentioning the results without deriving these.

$$E(\overline{x}) = \mu$$

and Var $(\bar{x}) = \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

i.e.$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

By comparing these expressions with the ones derived above we find that the standard error of is the same but further multiplied by a factor $\sqrt{(N-n)/(N-1)}$. This factor is, therefore, known as the finite population multiplier.

In practice, almost all the samples used picked up without replacement. Also, most populations are finite although they may be very large and so the standard error of the mean should theoretically be found by using the expression given above. However, if the population size (N) is large and consequently the sampling ratio (n/N) small, then the finite population multiplier is close to 1 and is not used, thus treating large finite populations as if they were infinitely large. For example, if N = 100,000 and n =100, the finite population multiplier

$$\sqrt{\frac{N-n}{N-1}} = \frac{100{,}000 - 100}{100{,}000 - 1}$$

$$= \frac{99{,}900}{99{,}999}$$

$$= .9995$$

Which is very close to 1 and the standard error of the mean would, for all practical purposes, be the same whether the population is treated as finite or infinite. As a rule of that, the finite population multiplier may not be used if the sampling ratio (n/N) is smaller than 0.05.

### Sampling from Normal Populations

We have seen earlier that the normal distribution occurs very frequently among many natural phenomena. For example, heights or weights of individuals, the weights of filled-cans from an automatic machine, the hardness obtained by heat treatment, etc. are distributed normally.

We also know that the sum of two independent random variables will follow a normal distribution if each of the two random variables belongs to a normal population. The sample mean, as we have seen earlier is the sum of n random variables $x_1, x_2, \ldots \ldots x_n$ each divided by n. Now, if each of these random variables is from the same normal population, it is not difficult to see that x would also be distributed normally.

Let $x \sim N(\mu, \sigma^2)$ symbolically represent the fact that the random variable x is distributed normally with mean μ and variance $\sigma^2$. What we have said in the earlier paragraphs, amounts to the following:

If $x \sim N(\mu, \sigma^2)$

then it follows that $x \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

The normal distribution is a continuous distribution and so the population cannot be small and finite if it is distributed normally; that is why we have not used the finite population multiplier in the above expression. We shall now show by an example, how to make use of the above result.

Suppose the diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm?

Let x be a random variable representing the diameter of one component picked up at random.

We know that $x \sim N(10, .01)$

Therefore, it follows that $x \sim N\left(10, \frac{.01}{5}\right)$

i.e. x will be distributed normally with a mean of 10 and a variance which is only of the variance of the population, since the sample size is 5.

$$\Pr\{9.95 \leqslant \bar{x} \leqslant 10.05\} = 2 \times \Pr\{10 \leqslant \bar{x} \leqslant 10.05\}$$

$$= 2 \times \Pr\left\{\frac{10-\mu}{\sigma/\sqrt{n}} \leqslant \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{10.05-\mu}{\sigma/\sqrt{n}}\right\}$$

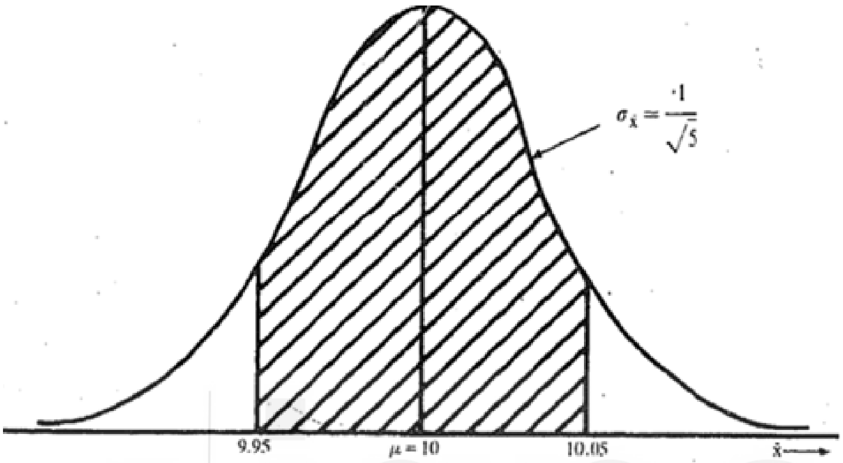
$$= 2 \times \Pr\left\{0 \leqslant z \leqslant \frac{10.05-10}{.1/\sqrt{5}}\right\}$$

$$= 2 \times \Pr\{0 \leqslant z \leqslant 1.12\}$$

$$= 2 \times 0.3686$$

$$= 0.7372$$

**Figure I: Distribution of x. The .Shaded area represents the probability that the random tunable x between 9.95 and 10.05.**



We first make use of the symmetry of the normal distribution and then calculate the z value by subtracting the mean and then dividing it by the standard deviation of the random variable distributed normally, viz k. The probability of interest is also shown as the shaded area in Figure I above.

## 10.3 CENTRAL LIMIT THEOREM

In this section we shall discuss one of the most important results of applied statistics which is also known by the name of the central limit theorem.

If $x_1, x_2, \ldots, x_n$ are n random variables which are independent and having the same distribution with mean µ. and standard deviation $\sigma$, then if $n \to \infty$, the limiting distribution of the standardised mean $z = \frac{x-\mu}{\sigma/\sqrt{n}}$ is the standard normal distribution.

In practice, if the sample size is sufficiently large, we need not know the population distribution because the central limit theorem assures us that the distribution of x can be approximated by a normal distribution. A sample size larger than 30 is generally considered to be large enough for this purposes.

Many practical samples are of size higher than 30. In all these cases, we know that the sampling distribution of the mean can be approximated by a normal distribution with an expected value equal to the population mean and a variance which is equal to the population variance divided by the sample size n.

We need to use the central limit theorem when the population distribution is either unknown or known to be non-normal. If the population distribution is known to be normal, then will also be distributed normally, as we have seen in section 14.2 above irrespective of the sample size.

**Activity A**

A sample of size 25 is picked up at random from a population which is normally distributed with a mean of 100 and a variance of 36. Calculate.

a) $\mathbf{P_r}\{\bar{x} \leqslant 99\}$

b) $\mathbf{P}_r\{98 \leqslant \bar{\mathbf{x}} \leqslant 100\}$

**Activity B**

If in (i) above, the sample is increased to 36, recalculate the following

a) $P_r\{\bar{x} \leqslant 99\}$

b) $P_r\{98 \leqslant \bar{x} \leqslant 100\}$

**Activity C**

Refer to Table 2 in the previous unit where we have a population of size 5.

A,B,C,D and E are five members of a family with the following weights of each family member:

$$\mathbf{x_a} = 70\text{kg}$$
$$x_b = 80\text{kg}$$
$$\mathbf{x_c} = 50\text{kg}$$
$$\mathbf{x_d} = 30\text{kg}$$
$$x_e = 10\text{kg}$$

Using the ten samples listed in Table 2, find the probability distribution of the sample mean and verify that

$$E(\bar{x}) = \mu = \frac{70 + 80 + 50 + 30 + 10}{5} = 48\text{kg}$$

$$\text{Var}\ (\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

## 10.4 SAMPLING DISTRIBUTION OF THE VARIANCE

We shall now discuss the sampling distribution of the variance. We shall first introduce the concept of sample variance and then present the chi-square distribution which helps us in working out probabilities for the sample variance, when the population is distributed normally.

**The Sample Variance**

By now it is implicitly clear that we use the sample mean to estimate the population mean, when that parameter is unknown. Similarly; we use a sample statistic called the sample variance to estimate the population variance. The sample variance is usually denoted by $s^2$ and it again captures some kind of an average of the square of deviations of the sample values from the sample mean. Let us put it in an equation form

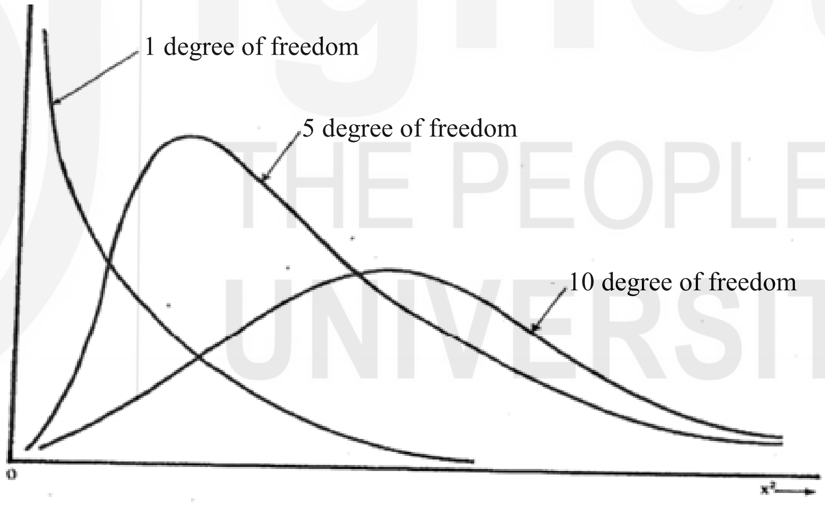
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

By comparing this expression with the corresponding expression for the population variance, we notice two differences. The deviations are measured from the sample mean and not from the population mean and secondly, the sum of squared deviations is divided by (n - 1) and not by n. Consequently, we can calculate the sample variance based only on the sample values without knowing the value of any population parameter. The division by (n - 1) is due to a technical reason to make the expected value of $s^2$ equal $\sigma^2$, which it is supposed to estimate.

**The Chi-square Distribution**

If the random variable x has the standard normal distribution, what would be the distribution of $x^2$? Intuitively speaking, it would be quite different from a normal distribution because now $x^2$, being a squared term, can assume only non-negative values. The probability density of $x^2$ will be the highest near 0, because most of the x value are close to 0 in a standard normal distribution. This distribution is called the chi-square distribution with 1 degree of freedom and is shown in Figure II below.

**Figure II: Chi-square ($x^2$) distribution with different degrees of freedom**



The chi-square distribution has only one parameter viz. the degrees of freedom and so there are many chi-square distributions each with its own degrees of freedom. In statistical tables, chi-square values for different areas under the right tail and the left tail of various chi-square distributions are tabulated.

If $X_1, X_2, \ldots, X_n$ are independent random variables, each having a standard normal distribution, then $(x_1^2 + x_2^2 + \cdots + x_n^2)$ will have a chi-square distribution with n degrees of freedom.

If $y_1$ and $y_2$ are independent random variables having chi-square distributions with $\gamma_1$ and $\gamma_2$ degrees of freedom, then $(y_1 + y_2)$ will have a chi-square distribution with $\gamma_1 + \gamma_2$ degrees of freedom.

We have stated some results above, without deriving them, to help us grasp the chi-square distribution intuitively. We shall state two more results in the same spirit.

If $y_1$ and $y_2$ are independent random variables such that $\gamma_1$ has a chi-square distribution with $\gamma_1$ degrees of freedom and $(y_1 + y_2)$ has a chi-square distribution with $\gamma > \gamma_1$ degrees of freedom, then $y_2$ will have a chi-square distribution with $(\gamma - \gamma_1)$ degrees of freedom.

Now, if $x_1, x_2, \ldots, x_n$ are n random variables from a normal population with mean $\mu$. and variance ..$\sigma^2$,

i.e.$x_i \sim N(\mu, \sigma^2), i = 1,2, \ldots, n$

it implies that $\frac{x_i - \mu}{\sigma} \approx N(0,1)$

and so $\left(\frac{x_1 - \mu}{\sigma}\right)^2$ will have a chi-square distribution with 1 degree of freedom.

Hence, $\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$ will have a chi-square distribution with n degrees of freedom.

We can break up this expression by measuring the deviation from x in place of .

We will then have

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{\sigma^2} \sum_{i-1}^{n} (\bar{x} - \mu)^2 + \frac{2(\bar{x} - \mu)}{\sigma^2} \sum_{i}^{n} (x_i - \bar{x})$$

$$= \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{x} - u}{\sigma/\sqrt{n}}\right)^2 \text{ since } \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Now, we know that the left hand side of the above equation is a random variable which has a chi-square distribution with n degrees of freedom. We also know that

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\therefore \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2$ will have a chi-square distribution with 1 degree of freedom. Hence, if the two terms on the right hand side of the above equation are independent (which will be assumed as true here and you will have to refer to advanced texts on statistics for the proof of the same), then it follows that $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with (n -1) degrees of freedom. One degree of freedom is lost because the deviations are measured from z and not from $\mu$.

### Expected Value and Variance of $s^2$

In practice, therefore, we work with the distribution of $\frac{(n-1)s^2}{\sigma^2}$ and not with the distribution of $s^2$ directly. The mean of a chi-square distribution is equal to its degrees of freedom and the variance is equal to twice the degrees of freedom. This can be used to find the expected value and the variance of $s^2$.

Since $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with (n–1) degrees of freedom,

$$\therefore E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1$$

or $\frac{(n-1)}{\sigma^2} \cdot E(s^2) = n-1$

$$\therefore E(s^2) = \sigma^2$$

Also Var $\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$

Using the definition of Variance, we get

$$E\left[\frac{(n-1)s^2}{\sigma^2} - E\left\{\frac{(n-1)s^2}{\sigma^2}\right\}\right]^2 = 2(n-1)$$

or, $E\left[\frac{(n-1)s^2}{\sigma^2} - (n-1)\right]^2 = 2(n-1)$

or, $\frac{(n-1)^2}{\sigma^4} \cdot E(s^2-\sigma^2)^2 = 2(n-1)$

$$\therefore E(s^2-\sigma^2)^2 = \frac{2\sigma^4}{n-1}$$

i.e.Var $(s^2) = \frac{2\sigma^4}{n-1}$

since the expected value of $s^2$ is equal to $\sigma^2$.

We therefore conclude that if we take a large number of samples, each with a sample size on n, from a normal population with mean μ and variance $\sigma^2$, each sample will perhaps have a different value for its sample variance $s^2$. But the average of a large number of values of $s^2$ will be close to $\sigma^2$. Also, the variance of $s^2$ falls as the sample size increases.

Let us recall that in all our discussion about the sampling distribution of the variance, we have been assuming that the population is distributed normally. If the population does not have a normal distribution, then nothing can be said about the distribution of $s^2$.

## 10.5 THE STUDENT'S DISTRIBUTION

We studied the sampling distribution of the mean in section 14.2 above where we showed that if the population distribution is normal then the distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is the standard normal distribution. In actual practice, the value of the population standard deviation $\sigma$ is often unknown which makes

ignou THE PEOPLE'S UNIVERSITY

it necessary to replace this with an estimate, usually by s-the sample standard deviation. In such cases, we would like to know the exact sampling distribution of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ for random samples from normal populations and this is provided by the t distribution which is also known as the student's t distribution after the pen name adopted by its author.

### The Concept of the t Statistic

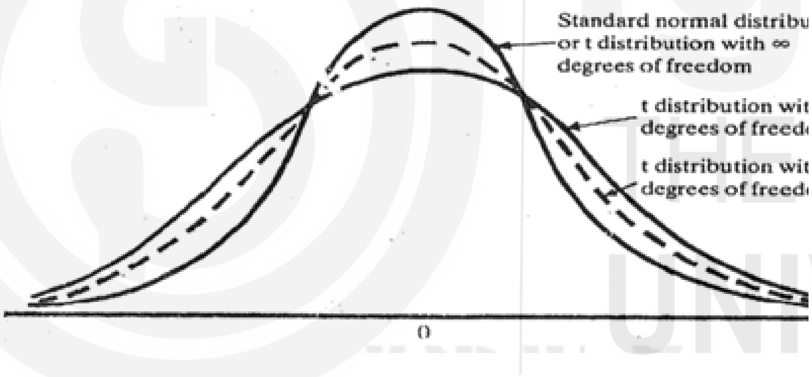
If x is a random variable having the standard normal distribution and y is a random variable having a chi-square distribution with v degrees of freedom and if x and y are independent, then the random variable

$$\mathbf{t} = \frac{\mathbf{x}}{\sqrt{\mathbf{y}/\mathbf{v}}}$$

has a distribution called the t distribution (or the Student's t distribution) with *v* degrees of freedom.

There are many t distributions, each with its degrees of freedom, which is the only parameter of this distribution. A t distribution is similar to the standard normal distribution as shown in Figure III below-only it is flatter and wider, thus having longer tails.

**Figure III: The t distribution with different degrees of freedom**



As the degrees of freedom increase, the t distribution comes closer to the standard normal distribution and when the degrees of freedom become infinitely large, the t distribution and the z distribution become indistinguishable.

### The t Distribution in Practice

If we have a random sample of size n from a normal population with mean $\mu$ and variance $\sigma^2$, then we know that the sample mean will be distributed normally with $\mu$ mean and variance $\sigma^2/n$. And so $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ will have a standard normal distribution.

We also know that in such a situation $\frac{(n-1)s^2}{\sigma^2}$ will have a chi-square distribution with (n -1) degrees of freedom. It has been shown in advanced texts that these two random variables are also independent and so $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)s^2}{\sigma^2(n-1)}}$ will have a t distribution with (n - 1) degrees of freedom.

After simplification, we conclude that $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ would have a t distribution with (n- 1) degrees of freedom.

It is therefore, possible to know the sampling distribution of x even when $\sigma$ is not known.

This result is really useful when the sample size is not very large. As we have seen earlier, if the sample size n is large, the t distribution with large degrees of freedom can be approximated by the z distribution. The t distribution is used when the degrees of freedom are not larger than 30; if the degrees of freedom are larger than 30, the t distribution is approximated by the standard normal or the z distribution.

The t distribution is again extensively tabulated because it is used quite frequently. As it is a symmetrical distribution, only one tail is generally tabulated and the other tail values can be worked out by using this property of symmetry.

## 10.6 SAMPLING DISTRIBUTION OF THE PROPORTION

Suppose we know that a proportion p of the population possesses a particular attribute that is of interest to use eg:- a proportion p of the population prefer our product to the next competing brand. This also implies that a proportion (1 - p) of the population do not prefer our product as compared to the next competing brand. If we pick up one member of the population at random, the probability of success i.e. the probability that this person will prefer our product to the next competing brand is p.

If the population is large enough, then even if we make repeated trials, which are considered to be independent, each with a probability of success equal to p. In such a case, if we make n repeated trials to pick up a sample of size n, the probability of x success in the sample is given by a binomial probability distribution, viz.

$$P(x) = \binom{\mathrm{n}}{\mathrm{x}} p^x (1-p)^{n-x}$$

If there are x successes in the sample, the sample proportion of success p is given by

$$\bar{\mathrm{p}} = \frac{\mathrm{x}}{\mathrm{n}}$$

The expected value and the variance of x, i.e. the number of successes in a sample of size n is known to be:

$$\bar{\mathrm{p}} = \frac{\mathbf{x}}{\mathbf{n}}$$

We can, therefore, find the expected value and the variance of the sample proportion p, as below:

$$E(\bar{\mathrm{p}}) = \mathrm{E}\left(\frac{\mathrm{x}}{\mathrm{n}}\right) = \frac{1}{\mathrm{n}}\mathrm{E}(\mathrm{x}) = \frac{1}{\mathrm{n}} \cdot \mathrm{np} = \mathrm{p}$$

and $\mathrm{Var}\,(\bar{p}) = \mathrm{Var}\left(\frac{x}{n}\right) = \frac{1}{n^2}\mathrm{Var}\,(x) = \frac{1}{n^2} \cdot np(1-p)$

$$= \frac{p(1-p)}{n}$$

Finally, if the sample size n is large enough, we can approximate the binomial probability distribution by a normal distribution with the same mean and variance. Thus, if n is sufficiently large,

$$\bar{\mathbf{p}} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

This approximation works quite well if n is sufficiently large so that both np and n(1- p) are at least as large as 5.

**Activity D**

A population is normally distributed with a mean of 100. A sample of size 15 is picked up at random from the population. If we know from t tables, that

$$P_r(t_{14} \geqslant 1.761) = 0.05$$

where $t_{14}$ represents a t variable with 14 degrees of freedom, calculate

$$P_r(\bar{x} \geqslant 115)$$

If we know that the sample standard deviation is 33.

Activity E

In a Board examination this year, 85% of the students who appeared for the examination passed. 100 students appeared in the same examination from School Q. What is the probability that 90 or more of these students passed?

……………………………………………………………………………….....

……………………………………………………………………………….....

……………………………………………………………………………….....

……………………………………………………………………………….....

……………………………………………………………………………….....

## 10.7 INTERVAL ESTIMATION

Suppose we want to estimate the mean income of a population of households residing in a part of a city. We might proceed by picking up a random sample of 100 households from the population and calculate the sample mean i.e. the mean income of the 100 sample households. In the absence of any other information, the sample mean can be .used as a point estimate of the population mean.

However, if we also want to convey the precision involved in this estimation, we need Distributions to give the standard error of the mean. As we have seen in section 14.2 above, the standard error of the mean depends on the population variance and the sample size.

The lower the standard error of the mean, the greater is the confidence on the correctness of our estimation. This process is further refined in interval estimation, wherein we present our estimate as an interval and quantify our confidence so that the true population parameter is contained by the estimated interval.

### The Confidence Level

As mentioned earlier, the sample mean is our estimate of the population mean. If we are asked to give an interval as our estimate, then we would add a range on the upper and the lower side of the sample mean and give that interval as our estimate. The larger the interval, the greater is our confidence that the interval does contain the true population mean. It is to be noted that the true population mean is a constant and is not a variable. On the other hand, the interval that we specify is a random interval whose position depends on the sample mean. For example if the sample mean is 50 and the standard error of the mean is 5, we may specify our interval estimate as (45,55) i.e. from 45 to 55 which spans one standard error of the mean on either side of the sample mean. On the other hand, if the interval estimate is specified as (40, 60) i.e. spanning two standard errors of the mean on either side of the sample mean, we are more confident that the latter interval contains the true population mean as compared to the former. However, if the confidence level is raised too high, the corresponding interval may become too wide to be of any practical use.

The confidence level, therefore, may be defined as the probability that the interval estimate will contain the true value of the population parameter that is being estimated. If we say that a 95% confidence interval for the population mean is obtained by spanning 1.96 times the standard error of the mean on either side of the sample mean, we mean that we take a large number of samples of size n, say 1000, and obtain the interval estimates from each of these 1000 samples and then 95% of these interval estimates would contain the true population mean.

### Confidence Interval for the Population Mean

We shall now discuss how to obtain a confidence interval for the population mean. We shill assume that the population distribution is normal and that the population aflame is known. Later, we shall relax the second condition.

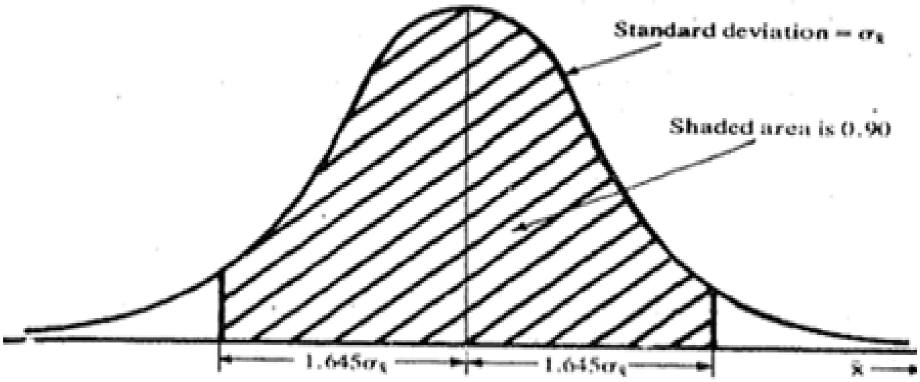
Suppose it is known that the weight of cement in packed bags is distributed normally with a standard deviation of 0.2 Kg. A sample of 25 bags is picked up at random and the mean weight of cement in these 25 bags is only 49.7 Kg. We want to find a 90% confidence interval for the mean weight of cement in filled bags.

Let x be a random variable representing the weight of cement in a bag picked up at random. We know that x is distributed normally with a standard deviation of 0.2 Kg.

The standard error of the mean can be easily calculated as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{25} = .04\text{Kg}$$

**Figure IV: Distribution of $\bar{x}$**



As shown in Figure IV above, we know that the sample mean is distributed normally with mean and standard deviation equal to 0.04 Kg. By referring to the normal table we can easily find that the probability that is between p. and $(\mu + 1.645\sigma_{\bar{x}})$ is 0.45 and so the probability that z is between $(\mu - 1.645\sigma_{\bar{x}})$ and $(\mu + 1.645\sigma_{\bar{x}})$ is 0.90. In other words, if we use an interval spanning from $(\bar{x} - 1.645\sigma_{\bar{x}})$ to $(\bar{x} + 1.645\sigma_{\bar{x}})$ then 90% of the time this interval will contain p,

Hence, for a 90% confidence interval,

the lower limit $= \tilde{x} - 1.645\sigma_{\bar{x}} = 49.7 - 1.645 \times 0.04$

$= 49.6342$

and the upper limit $= \bar{x} + 1.645\frac{\sigma}{x} = 49.7 + 1.645 \times 0.04$

$= 49.7658$

Therefore, we can state with 90% confidence level that the mean weight of cement in a filled hag lies between 49.6342 Kg and 49.7658 Kg.

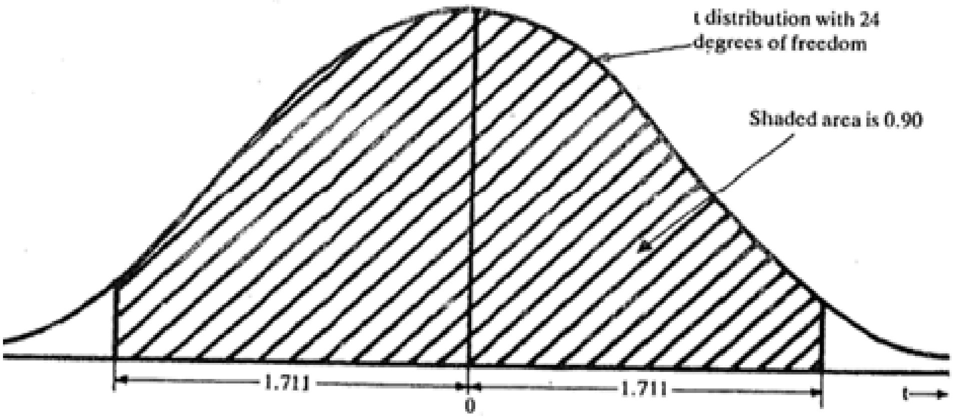
We can use the above approach when the population standard deviation is known or when the sample size is large n >30 , in which case the sample standard deviation can he used as an estimate of the population standard deviation. However, if the sample size is not large, as in the example above, then one has to use the t distribution in place of the standard normal distribution to calculate the probabilities.

Let us assume that we are interested in developing a 90% confidence interval in the same situation as described earlier with the difference that the population standard deviation is now not known. However, the sample standard deviation has been calculated and is known to be 0.2 Kg.

Since the sample size n = 25, we know that $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ follows a t distribution with degrees of freedom. From t tables, we can see that the probability that a t statistic with 24 degrees of freedom lying between $-1.711\ s/\sqrt{n}$ and 1.711

$s/\sqrt{n}$ is 0.90 – i.e. the probability that X lies between – 1.711 s/$\sqrt{n}$ and + 1.711 $s/\sqrt{n}$ is 0.90. This is shown in Figure 5 below.

**Figure V: Area under a t distribution**



In other words, if we use an interval spanning from ($\bar{x} - 1.711\, s/\sqrt{n}$ to ($\bar{x} + 1.711 s/\sqrt{n}$) then 90% of the time, this interval will contain $\mu$. Hence, for a 90% confidence interval,

the lower limit $= \bar{x} - 1.711 \frac{s}{\sqrt{n}} = 49.7 - 1.711 \frac{0.2}{\sqrt{25}}$

$= 49.6316$

and the upper limit $= \bar{x} + 1.711 \frac{s}{\sqrt{n}} = 49.7 + 1.711 \frac{0.2}{\sqrt{25}}$

$= 49.7684$

In this case, we can state with 90% confidence level that the mean weight of cement in a filled hag lies between 49.6316 Kg and 49.7684 Kg.

## 10.8 THE SAMPLE SIZE

In earlier section above we have seen how the sampling distribution of a statistic helps us in developing a confidence interval for the corresponding population parameter. In this section we shall present another application of the sampling distributions. We have earlier referred to the fact that in some situations the sample size required can he determined on the basis of the precision of the estimates. We shall now demonstrate this process.

**Sample Size for Estimating Population Mean**

We assume that the population distribution is normal and the population standard deviation is known. In such a case the sample size required for a given confidence level and a required accuracy can he easily determined. We again take the help of an example.

Suppose we know that the weight of cement in filled bags is distributed normally with a standard deviation o of 0.2 Kg. We want to know how large a sample should he taken so that the mean weight of cement in a filled hag can be estimated within plus or minus 0.05 Kg of the true value with a confidence level of 90%.

We have seen in section 14.7 above that the interval $\left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}\right)$ to $\left(\bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right)$ contains the true value of the population mean 90% of the time. We also want that the interval (X-0.05) to (X+0.05) should give us a 90% confidence level.

Therefore, $1.645\frac{\sigma}{\sqrt{n}} = 0.05$

And so n = $\left(\frac{1.645\times0.2}{0.05}\right)^2$

= 43.3

We must have a sample size of at least 44 so that the mean weight of cement in a filled bag can be estimated within plus or minus 0.05 Kg of the true value with a 90% confidence level.

It is to be noted that this approach does not work if the population standard deviation is not known because the sample standard deviation is known only after the sample has been analysed whereas the sample size decision is required before the sample is picked up.

**Sample Size for Estimating Population Proportion**

Suppose we want to estimate the proportion of consumers in the population who prefer our product to the next competing brand. How large a sample should be taken so that the population proportion can be estimated within plus or minus 0.05 with a 90% confidence level?

We shall use the sample proportion $\bar{p}$ to estimate the population proportion p. If n is sufficiently large, the distribution of $\bar{p}$ can be approximated by a normal distribution with mean p and variance p (1 - p)/n (let q = 1 – p).

From normal tables, we can now say that the probability that p will lie between $(\mathrm{p} - 1.645\sqrt{\mathrm{pq/n}}$ and $(\mathrm{p} + 1.645\sqrt{\mathrm{pq/n}}$ is 0.90. In other words, the interval $(\bar{\mathrm{p}} - 1.645\sqrt{\mathrm{pq/n}}$ to $(\bar{\mathrm{p}} + 1.645\sqrt{\mathrm{pq/n}}$ will contain p, 90% of the time.

We also want that the interval (p - 0.05) to (p + 0.05) should contain p, 90% of the time.

Therefore, $1.645\sqrt{\frac{\mathrm{p(1-p)}}{\mathrm{n}}} = 0.05$

or $\sqrt{\frac{p(1-p)}{n}} = \frac{0.05}{1.645} = 0.0304$

or $\frac{p(1-p)}{n} = 0.0009239$

$$\therefore n = \frac{p(1-p)}{0.0009239}$$

But we do not know the value of p, so n cannot be calculated directly. However, whatever be the value of p, the highest value for the expression p (1 - p) is 0.25, which is the case when p = 0.5. Hence, in the worst case the highest possible value for p(1 -p) is 0.25. In that case 0.25

$$n = \frac{0.25}{0.0009239} = 270.6$$

Therefore, if we take a sample of size 271, then we are sure that our estimate of the population proportion would be within plus and minus 0.05 of the true value with a confidence level of 90% whatever he the value of p.

**Activity F**

100 Sodium Vapour Lamps were tested to estimate the life of such a lamp. The life of these 100 lamps exhibited a mean of 10,000 hours with a standard deviation of 500 hours. Construct a 90% confidence interval for the true mean life of a Sodium Vapour Lamp.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity G**

If the sample size in the previous situation had been 15 in place 100, what would be the confidence interval.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

**Activity H**

We want to estimate the proportion of employees who prefer the codification of rules and regulations. What should be the sample size if we want our estimate to he within plus or minus 0.05 with a 95% confidence level.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

## 10.9 SUMMARY

We have introduced the concept of sampling distributions in this unit. We have discussed the sampling distributions of some commonly used statistics and also shown some applications of the same.

A sampling distribution of a sample statistic has been introduced as the probability distribution or the probability density function of the sample statistic. In the sampling distribution of the mean, we find that if the population distribution is normal, the sample mean is also distributed

ignou THE PEOPLE'S UNIVERSITY

normally with the same mean but with a smaller standard deviation. In fact, the standard deviation of the sample mean, also known as the standard error of the mean, is found to be equal to the population standard deviation divided by the sample size.

We have also presented a very important result called the central limit theorem which assures us that if the sample size is large enough (greater than 30), the sampling distribution of the mean could be approximated by a corresponding normal distribution with the mean and standard deviation as given in the preceding paragraph.

We have then explored the sampling distribution of the variance and found that a related quantity viz. $\frac{(n-1)s^2}{\sigma^2}$ would have a chi-square distribution with (n -1) degrees of freedom. We have learnt that the chi-square distribution is tabulated extensively and so any probability calculations regarding $s^2$ could be easily made by referring to the tables for the chi-square distribution.

We have introduced one more distribution viz. the t distribution which is found to be applicable when the sampling distribution of the mean is of interest, but the population standard deviation is unknown. It is noticed that if the sample size is large enough (n>30), the t distribution is actually very close to the standard normal distribution.

We have also studied the sampling distribution of the proportion and then looked at two applications of the sampling distributions. One is in developing an interval estimate for a population parameter with a given confidence level, which is conceptualised as the probability that a random interval will contain the true value of the parameter. The second application is to determine the sample size required while estimating the population mean or the population proportion.

## 10.10 SELF-ASSESSMENT EXERCISES

1) What is the practical utility of the central limit theorem in applied statistics?

2) The daily wages of a random sample of farm labourers are:

   14 17 14.5 22 27 16.5 .19.5 21 18 22.5

   i) What is the best estimate of the mean daily wages of all farm labourers?

   ii) What is the standard error of the mean?

   iii) What is the 95% confidence interval for the population mean? Explain what it indicates and also any assumption you made before you could calculate the confidence interval.

3) An inspector wants to estimate the weight of detergent in packets filled by an automatic filling machine. She wants to be 95% confident that her estimate is not away from the true mean weight of detergent by more than 10 gms. What should the minimum sample size be if it is known that the

standard deviation of the weight of detergent filled by that machine is 100 gms?

4) A steamer is certified to carry a load of 20,000 Kg. The weight of one person is distributed normally with a mean of 60 Kg and a standard deviation of 15 Kg.

   i) What is the probability of exceeding the certified load if the steamer is carrying 340 persons?

   ii) What is the maximum number of persons that can travel by the steamer at any time if the probability of exceeding the certified load should not exceed 5%?

   Indicate the most appropriate choice for each of the following situations:

5) The finite population multiplier is not used when dealing with large finite population because

   i) when the population is large, the standard error of the mean approaches zero.

   ii) another formula is more appropriate in such cases.

   iii) the finite population multiplier approaches 1.

   iv) none of the above.

6) When sampling from a large population, if we want the standard error of the mean to be less than one-half the standard deviation of the population, how large would the sample have to be?

   i) 3

   ii) 5

   iii) 4

   iv) none of these

7) A sampling ratio of 0.10 was used in a sample survey when the population size was 50. What should the finite population multiplier be?

   i) 0.958

   ii) 0.10

   iii) 1.10

   iv) cannot be calculated from the given data.

8) As the sample size is increased, the standard error of the mean would

   i) increase in magnitude

   ii) decrease in magnitude

   iii) remain unaltered

   iv) may either increase or decrease.

9) As the confidence level for a confidence interval increases, the width of the interval

ignou THE PEOPLE'S UNIVERSITY

i) Increases

ii) Decreases

iii) remains unaltered

iv) may either increase or decrease.

## 10.11 FURTHER READINGS

Emory, L.W., *Business Research Methods, Richard* D. Irwin, Inc: Homewood.

Ferber, R.(ed.), *Handbook of Marketing Research,* McGraw Hill Book Co.: New York.

Levin, R.I., *Statistics for Management, Prentice* Hall of India: New Delhi.

Mason, R.D., *Statistical Techniques in Business and Economics*, Richard D. Irwin, Inc: Homewood.

Mendenhall, W., R.L. Scheaffer and D.D. Wackerly, *Mathematical Statistics with Applications*, Dunbury Press: Boston.

Plane, D.R. and E.B. Oppermann, *Business and Economic Statistics*, Business Publications, Inc: Plano.

Bruce Bowernman. Business statistics for practices; McGraw Hill.

ignou THE PEOPLE'S UNIVERSITY

# UNIT 11 TESTING OF HYPOTHESES

**Objectives**

Upon successful completion of the unit, you should be able to:

- understand the meaning of statistical hypothesis
- absorb the concept of the null hypothesis
- appreciate the importance of the significance level and the P value of a test
- learn the steps involved in conducting a test of hypothesis
- perform tests concerning population mean, population proportion, difference between the population means and two population proportions.

**Structure**

11.1 Introductions
11.2 Some Basic Concepts
11.3 Hypothesis Testing Procedure
11.4 Testing of Population Mean
11.5 Testing of Population Proportion
11.6 Testing for Differences Between Means
11.7 Testing for Differences Between Proportions
11.8 Summary
11.9 Self-assessment Exercises
11.10 Further Readings

ignou
THE PEOPLE'S
UNIVERSITY

## 11.1 INTRODUCTION

In this unit and the next, we shall study a class of problems where the decision made by a decision maker depends primarily on the strength of the evidence thrown up by a random sample drawn from a population. We can elaborate this by an example where the purchase manager of a machine tool making company has to decide whether to buy castings from a new supplier or not. The new supplier claims that his castings have higher hardness than those of the competitors If the claim is true, then it would be in the interest of the company to switch from the existing suppliers to the new supplier because of the higher hardness, all other conditions being similar. However, if the claim is not true, the purchase manager should continue to buy from the existing suppliers. He needs a tool which allows him to test such a claim.

Testing of hypothesis provides such a tool to the decision maker. If the purchase manager were to use this tool, he would ask the new supplier to deliver a small number of castings. The sample of castings will be evaluated and based on the strength of the evidence produced by the sample, the purchase manager will accept or reject the claim of the new supplier and accordingly make his decision. The claim made by the new supplier is a hypothesis that needs to be tested and a statistical procedure which allows us to perform such a test is called testing of hypothesis.

**What is a Hypothesis**

A hypothesis, or more specifically a statistical hypothesis, is some statement about a population parameter or about a population distribution. If the population is large, there is no way of analysing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of a random sample.

Our hypothesis for the example situation in 11.1 could be that the mean hardness of castings supplied by the new supplier is less than or equal to 20, where 20 is the mean hardness of castings supplied by existing suppliers.

**A Two-action Decision Problems**

The decision problem faced by the purchase manager in 11.1 above has only two alternative courses of action-either to buy from the new supplier or not to buy from the new supplier. The alternative chosen depends on whether the claim made by the new supplier is accepted or rejected. Now, the claim made by the new supplier can be formulated as a statistical hypothesis-as has been done in 11.1 above. Therefore, the decision made or the alternative chosen depends primarily on whether a hypothesis is accepted or rejected.

ignou
THE PEOPLE'S
UNIVERSITY

## 11.2 SOME BASIC CONCEPTS

We shall now discuss some concepts which will come in handy when we attempt to set up a procedure for testing of hypothesis.

**The Null Hypothesis**

As stated earlier, a hypothesis is a statement about a population parameter or about a population distribution. In any testing of hypothesis problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one the alternative hypothesis. The null hypothesis is represented as $H_0$ and the alternative hypothesis is represented as $H_I$. For example, if the population mean is represented by we can set up our hypothesis , as follows:

$$H_0: \mu = 20$$

$$\mathrm{H}_1: \mu > 20$$

What we have represented symbolically above can be interpreted to mean that the null hypothesis is that the population mean is not greater than 20, whereas the alternative hypothesis is that the population mean is greater than 20. It is clear that both $H_0$ and $H_I$ cannot be true and also that one of them will always be true. At the end of our testing procedure, if we come to the conclusion that $H_0$ should be, rejected, this also amounts to saying that $H_I$ should be accepted and vice versa.

It is not difficult to identify the pair of hypotheses relevant in any decision situation. Can any one of the two be called the' null hypothesis? The answer is a big no-because the roles of $H_0$ and $H_1$ are not symmetrical.

One can conceptualise the whole procedure of testing of hypothesis as trying to answer one basic question: Is the sample evidence strong enough to enable us to reject $H_0$? This means that $H_0$ will be rejected only when there is strong sample evidence against it. However, if the sample evidence is not strong enough, we shall conclude that we cannot reject $H_0$ and so we accept $H_0$ by default. Thus, $H_0$ is accepted even without any evidence in support of it whereas it can be rejected only when there is an overwhelming evidence against it.

Perhaps the problem faced by the purchase manager in above will help us in understanding the role of the null hypothesis better. The new supplier has claimed that his castings have higher hardness than the competitor's. The mean hardness of casting supplied by the existing suppliers is 20 and so the purchase manager can test the claim of the new supplier by setting up the following hypotheses:

$$H_0; \mu \leqslant 20$$

$$H_1: \mu > 20$$

In such a case, the purchase manager will reject the null hypothesis only when the sample evidence is overwhelmingly against it-e.g. if the sample mean from the sample of castings supplied by the new supplier is 30 or 40, this evidence might be taken to be overwhelmingly strong so that Ho can be rejected and so purchase effected from the new supplier. On the other hand if the sample mean is 20.1 or 20.2, this evidence may be found to be too mild to reject $H_a$ so that $H_0$ is accepted even when the sample evidence is against it.

In other words, the decision maker is somewhat biased towards the null hypothesis and he does not mind accepting the null hypothesis. However, he would reject the null hypothesis only when the sample evidence against the null hypothesis is too large to be ignored. We shall explore the reasons for this bias below.

The null hypothesis is called by this name because in many situations, acceptance of this hypothesis would lead to null action. For example, if our purchase manager accepts the null hypothesis, he would continue to buy castings from the existing suppliers and so status quo ante would be maintained. On the other hand, rejecting the null hypothesis would lead to a change in status quo ante and purchase is to be made from the new supplier.

### Type I and Type II Errors

Since we are basing our conclusion on the evidence produced, by a sample and since variations from one sample to another can never be eliminated until the sample is as large as the population itself, it is possible that the conclusion drawn is incorrect which leads to an error. As shown in Table 1 below, there can be two types of errors and for convenience, each of these errors have been given a name.

**Table 1: Types of Errors in Testing of Hypothesis**

| | | Decision based on Sample | |
|---|---|---|---|
| | | Accept $H_0$ (Reject $H_1$) | Reject $H_0$ (Accept $H_1$) |
| States of Population | $H_0$ is True ($H_1$ is False) | Correct | Wrong (Type I Error) |
| | $H_0$ is False ($H_1$ is True) | Wrong (Type II Error) | Correct |

If we wrongly reject Ho , when in reality Ho is True-the error is called a type I error. Similarly, when we wrongly accept Ho when Ho is False--the error is called a type II error. Let us go back to the decision problem faced by the purchase manager, referred to in the Null Hypothesis above. If the purchase manager rejects $H_0$ and places orders with the new supplier when the mean hardness of the castings supplied by the new supplier is in reality no better than the mean hardness of castings supplied by the existing suppliers, he would be making a type I error. In this situation, a type II error would mean not to buy castings from the new supplier when his castings are really better.

Both these errors are bad and should be reduced to the minimum. However, they can be completely eliminated only when the full population is examined-in which case there would be no practical utility of the testing procedure. On the other hand, for a given sample size, these two errors neutralise each other as we shall see later in this unit. This implies that if the testing procedure is designed as to reduce the probability of occurrence of type I error, simultaneously the probability of type II error would go up and vice versa. What can at best be achievedr is a reasonable balance between these two errors.

In all testing of hypothesis procedures, it is implicitly assumed that type I error is much more severe than type II error and so needs to be controlled. If we go back to the purchase manager's problem, we shall notice that type I error would result in a real financial loss to the company since the company would have switched from the existing suppliers to the new supplier who is in reality no better. The new castings are no better and perhaps worse than the earlier ones thus affecting the quality of the final product (machine tools) produced. On top of it, the new supplier might be given a higher rate for his castings as these have been claimed to have higher hardness. And finally, there is a cost associated with any change.

Compared to this, type II error in this situation would result to an opportunity loss since the company would forego the opportunity of using better castings. The greater the difference in costs between type I and type II errors, the stronger would be the evidence needed to be able to reject $H_0$-i.e. the probability of type I error would be kept down to lower limits. It is to be noted that type I error occurs only when $H_0$ is wrognly rejected.

ignou
THE PEOPLE'S
UNIVERSITY

## The Significance Level

In all tests of hypothesis, type I error is assumed to be more serious than type II error and so the probability of type I error needs to be explicitly controlled. This is done through specifying a significance level at which a test is conducted. The significance level, therefore, sets a limit to the probability of type I error and test procedures are designed so as to get the lowest probability of type II error subject to the significance level.

The probability of type I error is usually represented by the symbol $\alpha$ (read as alpha) and the probability of type II error represented by $\beta$ (read as beta).

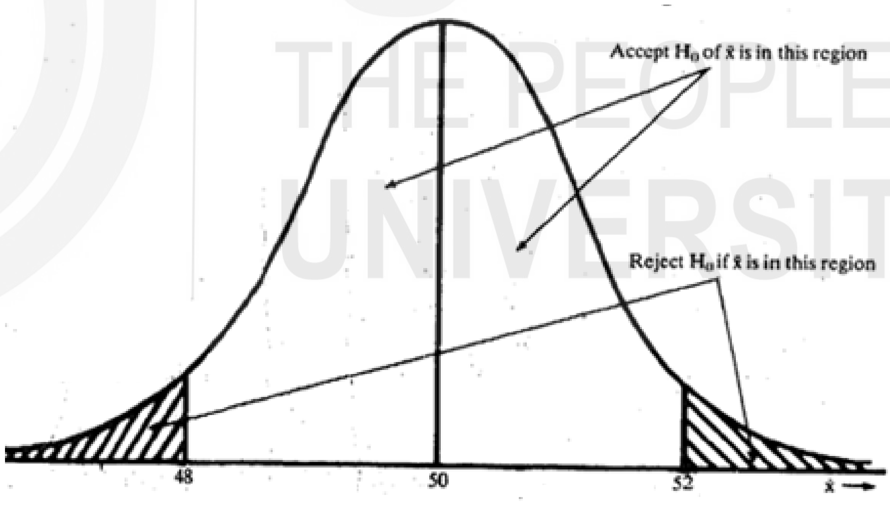
Suppose we have set up our hypotheses as follows:

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

We would perhaps use the sample mean $\bar{x}$ to draw inferences about the population mean $\mu$. Also, since we are biased towards Ho we would be compelled to reject Ho only when the sample evidence is strongly against it. For example, we might decide to reject Ho only when > 52 or x<48 and in all other cases i.e. when x is between 48 and 52 and so is close to 50, we might conclude that the sample evidence is not strong enough for us to be able to reject Ho.

**Figure I: The significance Level h the area of the shaded region**



Now suppose the $H_0$ is in reality true--i.e. the true value of p is 50. In that case, if the population distribution is normal or if the sample size is sufficiently large (n > 30), the distribution of z will be normal as shown in Figure I above. Remember that our criterion for rejecting Ho states that if x< 48 or x> 52, we shall reject Ho. Referring to Figure I, we find that the shaded area (under both tails 'of the distribution of t represents the probability of rejecting $H_0$ when $H_0$ is true which is the same as the probability of type I error.

All tests of hypotheses hinge upon this concept of the significance level and it is possible that a null hypothesis can be rejected at $\alpha = .05$ whereas the same evidence is not strong enough to reject the null hypothesis at $\alpha = .01$. In

ignou
THE PEOPLE'S
UNIVERSITY

other words, the inference drawn can be sensitive to the significance level used.

Testing of 'hypothesis suffers, from the limitation that the financial or the economic costs of consequences are not considered explicitly. In practice, the significance level is supposed to be arrived at after considering the cost consequences. It is very difficult to specify the ideal value of a in a specific situation; we can only give a guideline that the higher the difference in costs between type I error and type II error, the greater is the importance of type I error as compared to type II error. Consequently, the risk or probability of type I error should be lower-i.e. the value of should be lower. In practice, most tests are conducted at $\alpha = .01$, $\alpha = .05$ or $\alpha = .1$ by convention as well as by convenience.

### The Power Curve of a Test

Let us go back to the purchase manager's problem referred to earlier where we set up our hypotheses as follows:

$$\mathbf{H}_0 : \mu \leqslant$$

$$\mathbf{H}_1 : \mu > 20$$

These hypotheses imply that the purchase manager would normally accept the null hypothesis that the mean hardness of castings delivered by the new supplier is not above 20-in which case no purchase order need be placed with the new supplier. Only when the sample evidence is strongly against it, would the null hypothesis be rejected-in which case the purchase manager would place orders with the new supplier.

Now suppose that the purchase manager knows that the hardness of castings from any supplier is normally distributed and also that the standard deviation of hardness of castings from the new supplier would not be much different from that of the existing suppliers which is known to be 2.5. Further, suppose the purchase manager picks up a sample of 100 castings and he decides that if the sample mean from these 100 castings is greater than or equal to 20.5, he would consider the sample evidence to be strongly against Ho and so he would reject Ho. The test is now completely designed and has been summarised as follows:

$$\mathrm{H}_0 : \mu \leqslant 20$$

$$\mathrm{H}_1 : \mu > 20$$

Reject $\mathrm{H}_0$ if $\bar{x} \geqslant 20.5$ for n = 100, where $\sigma = 2.5$

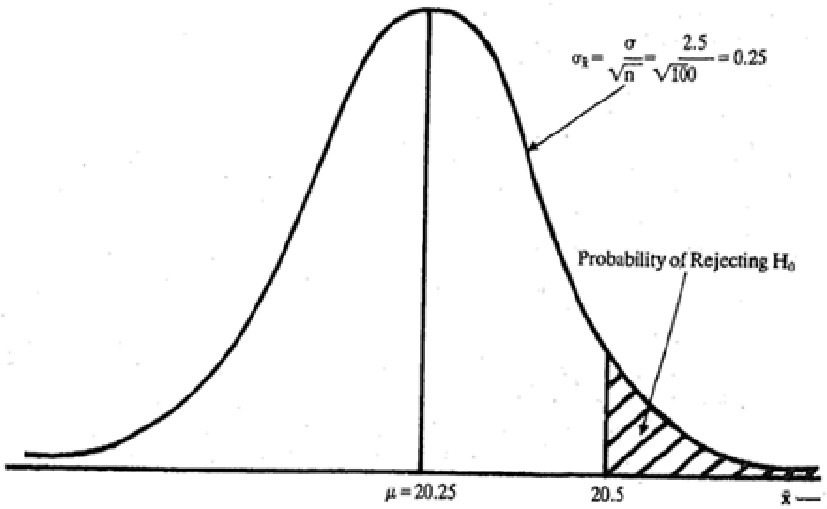
For this test, we can easily calculate the probability that Ho would be rejected for a given value of $\mu$. For example, if we know that the true value of $\mu$ is 20.25, the probability that Ho is rejected is given by the shaded area in Figure II below.

$$\Pr\,[\bar{x} \geqslant 20.5] = \Pr\left[\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geqslant \frac{20.5 - 20.25}{2.5/\sqrt{100}}\right]$$

$$= \Pr\,[z \geqslant 1]$$

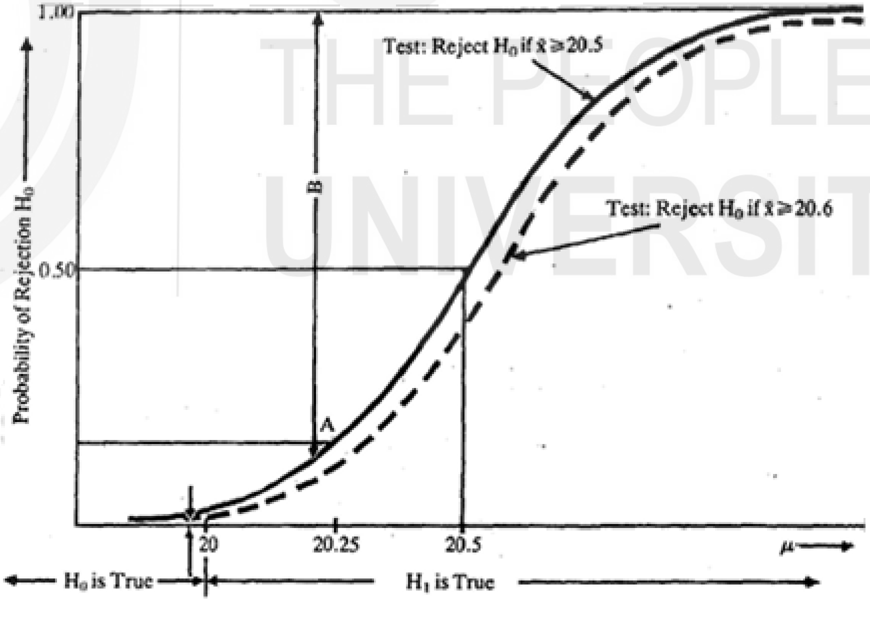
$$= 0.1587$$

**Figure II: Probability of rejecting Ho when $\mu$ = 20.25**



We can similarly calculate the probability of rejecting Ho for different values of p, and plot these on a graph as shown in Figure III below. Such a curve is known as the Power Curve of a test. Point A on this power curve, for example, can be interpreted to mean that if $\mu = 20.25$, then the probability of rejecting Ho is 0.1587. Incidentally, this is the probability that we calculated in the previous paragraph.

**Figure III: Power curve of a Test.**



We have also marked two regions-one where Ho is true ($\mu \leq 20$) and the other where $H_1$ is true ($\mu > 20$). We have also marked A for one value of 20 and similarly marked B for another value of $\mu > 20$. The dotted line shows the power curve of another test [Reject Ho if $\bar{x} \geq 20.6$] conducted on a sample of the same size. By comparing the power curve of these two tests we see very clearly that for a given sample size, a reduces as β increases and vice versa.

We also see in Figure III that in the range where Ho is true viz $\mu \leq 20$, the value of α is different for different values of $\mu$-but the highest value of a occurs at the breakpoint between Ho and $H_1$ i.e at $\mu = 20$. In other words, the probability of type I error is highest when $\mu = 20$, which is the breakpoint value between Ho and $H_1$ Therefore, if we want to ensure that the probability of type I error does not exceed a particular value (say 0.05), it is enough to check that the probability of type I error does not exceed this value at the breakpoint value of $\mu$. This property will be used very frequently in designing the tests. It is to be noted that when we specified the test as: Reject Ho if $x \geq 20.5$, we partitioned all possible values of x into two regions-one can be called the acceptance region (viz.20.5) and the other the rejection region or the critical region (viz.20.5). If the value of the sample statistic lies in the critical region, then only can we reject Ho.

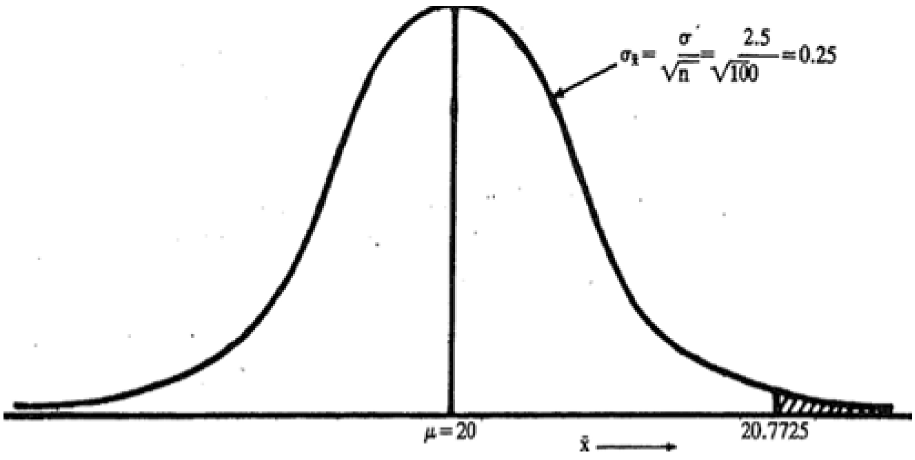
### The P Value of a Test

We have seen earlier that a test of hypothesis is designed for a significance level and at the end of the test we conclude that we reject the null hypothesis at 1% significance level and so on. As discussed earlier, the significance level is somewhat arbitrarily fixed and the mere fact that a hypothesis is rejected or cannot be rejected does not reveal the full strength of the sample evidence. An alternative, and in some ways, a better way of expressing the conclusion of a test is to state the P value or the probability value of the test.

The P value of a test expresses the probability of observing sample statistic as extreme as the one observed if the null hypothesis is true. We shall use the purchase manager's decision problem discussed above, under the subheading The Power Curve of a Test, to explain the P value. Please go through that section before you proceed further.

Suppose the observed value of the sample mean k, from a sample of size 100, is 20.7725. What is the significance level at which we shall just reject Ho? Or in other words, what is the probability of observing an x of 20.7725 when Ho is true? We now know that the probability of type I error is the highest when the population parameter is at the breakpoint value between Ho and Hi and so the highest probability of type I error occurs if we reject the null hypothesis when $\bar{x}$= 20.7725 and $\mu = 20$ and this probability can be calculated as shown in Figure IV below.

**Figure IV: The P value of a Test**



$$\Pr\left[\bar{x} \geqslant 20.7725\right] = \Pr\left[\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \geqslant \frac{20.7725-20}{2.5/\sqrt{100}}\right]$$

$$= \Pr\left[z \geqslant 3.09\right]$$

$$=0.001$$

Thus, we can say that the P value of this test is 0.001 and this is more meaningful to say than that we reject the null hypothesis at $\alpha = 0.05$ or at $\alpha = 0.01$

## 11.3 HYPOTHESIS TESTING PROCEDURE

By now it should be clear that there are basically two phases in testing of hypothesis¬in the first phase we design the test and set up the conditions under which we shall reject the null hypothesis. In the second phase we use the test based on the sample evidence and draw our conclusion as to whether the null hypothesis can be rejected (or else, what is the P value of the test). The detailed steps involved are as follows:

Step 1: State the Null and the Alternate Hypotheses.

Step 2: Choose the test statistic-i.e. the sample statistic that will define the critical region.

Step 3: Specify a level of significance of $\alpha$.

Step 4: Define the critical region in terms of the test statistic.

Step 5: Compare the observed value of the test statistic with the cut-off value or the critical value and decide to accept or reject the null hypothesis.

The best way to explain these steps is through an example and that is what we propose to do forthwith.

**Activity A**

Is it possible that a false hypothesis will be accepted? Does it mean that we are never sure of our conclusion?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

**Activity B**

Suppose we are testing the mean of a population and the test procedure is: Reject Ho if $\bar{x}$= 25.5, If the standard error of the mean is known to be 0.5 then calculate the probability of accepting Ho when in reality it is not true and $\mu$ = 25. Should we use a or 3 to represent this probability?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

ignou
THE PEOPLE'S
UNIVERSITY

**Activity C**

Name one situation from your work where you think testing of hypotheses might be of use to you.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

## 11.4 TESTING OF POPULATION MEAN

We shall now discuss how tests concerning population means can be developed and used. Under different conditions, the test procedures have to be developed differently. We start by discussing the case when the population variance is known and the distribution of sample mean z is known to have or can be approximated by a normal distribution.

**When Population Variance is Known**

We again refer to the purchase manager's decision problem first introduced in section 11.1 and elaborated again in 11.2. The purchase manager has to decide whether to buy castings from a new supplier who has claimed that his castings have higher hardness than those supplied by existing suppliers. The purchase manager knows that the mean hardness of castings supplied by existing suppliers is 20 and also that the standard deviation of hardness is 2.5. To test the claim of the new supplier, he picks up a sample of 100 castings from the new supplier and finds that the sample mean is 20.5. The purchase manager believes that the standard deviation of hardness of castings from the new supplier would not be very different from that of the existing suppliers. If the purchase manager decides to use a significance level of 5%, what should we conclude?

We have seen earlier that unless and until the sample evidence is strongly to the contrary, the purchase manager would not like to switch from the existing suppliers. The null and the alternative hypotheses are, therefore, set up as follows:

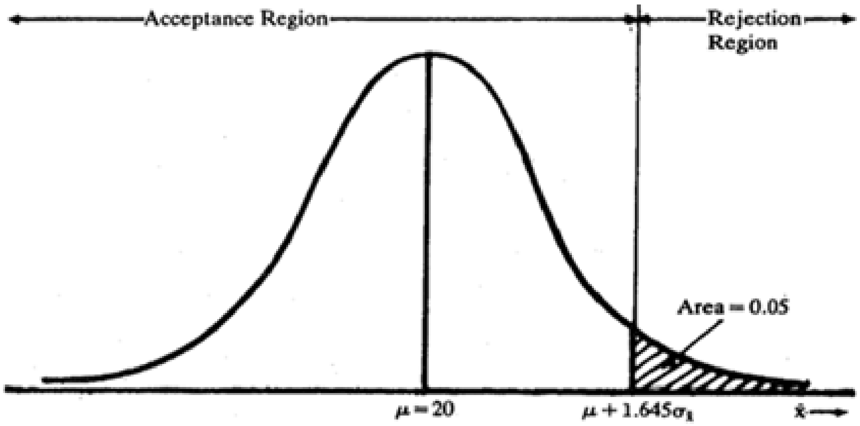
$$\mathrm{H_o}{:}\times \mu \leqslant 20$$

$$H_1{:}\, \mu > 20$$

The sample mean would be used to draw conclusions about the population mean and so the test statistic is Z. We shall be in a position to reject Ho only if the sample evidence is strongly against it i.e. if the observed value of *x* is much larger than 20. The critical region will therefore be of the form: x≤c, where c is a real number much larger than 20. The actual value of c would depend on the significance level used.

The significance level is known to be, α = 0.05. In other words, the probability of type I error should not exceed 0.05. We also know that the probability of type I error is highest when p, is at the breakpoint value between Ho and $H_1$ i.e. when $\mu = 20$.

**Figure V: The Critical Region**



This has been shown as the shaded region in Figure V above, where the distribution of has been shown as a normal curve. This is valid under two conditions-(1) if the population distribution is normal, then the distribution of z is also normal, or (2) if the 'sample size is large, then again, the central limit theorem assures us that the distribution of x can be approximated by a normal distribution. Therefore, if either of these conditions is valid (and in this case the second condition is certainly valid as n = 100), then

$$\tilde{x} \sim N(\mu, v_{\tilde{x}}^2)$$

where the population mean is $\mu$ and $\sigma_{\bar{x}}$) is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{100}}$$

=0.25

As shown in Figure V, we want a value of $\bar{\mathbf{x}}$ such that the area to the right of this value is 0.05 when $\mu = 20$. By referring to normal tables, we can find

the cut-off value of $\bar{\mathbf{x}} = \mu + 1.645\sigma_{\bar{\mathbf{x}}}$, When $\mu = 20$

= 20+1.645x0.25

= 20.41125

Hence, the test procedure boils down to:

Reject $H_0$ if $\bar{x} \geqslant 20.41125$

Now that we have identified the critical region, we can compare the observed value of x and see if it belongs to critical region. The observed value of x is 20.5-which lies in the critical region and so we can conclude that the sample evidence is strong enough for us to reject Ho.

### One-tailed and Two-tailed Tests

In the previous section we looked at a test where the critical region was found to lie under one tail-the right tail-of the distribution of the test statistic. Such tests are called one-tailed tests in contrast with the two-tailed tests where the critical region lies under both the tails of the distribution of the test statistic. We shall now look at such a situation.

Let us assume that our purchase manager wants to test whether the mean hardness of castings supplied by one of the existing suppliers has changed from 20. If it has changed from 20, then he would like to take some corrective action. On the other hand, he would not like to initiate the corrective actions unless and until he is reasonably sure that the mean hardness has really changed. So, he tests a sample of 49 castings from this supplier and finds that the mean hardness is 19.5. What should he conclude at a significance level ($\alpha$) of 0.05? Assume that a continues to be 2.5.

To begin with, we state our hypotheses as

$$H_0: \mu = 20$$

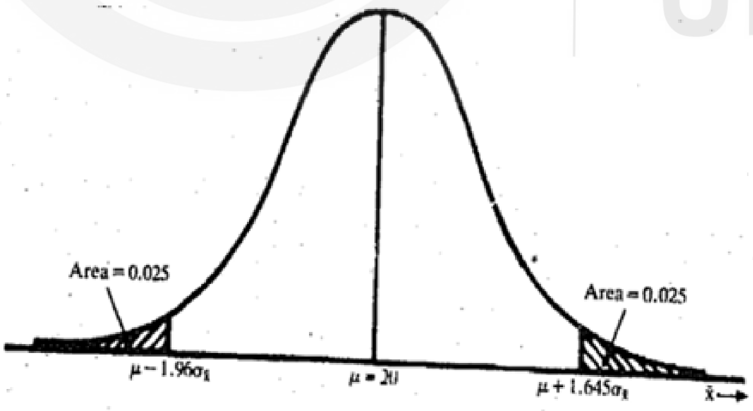
$$H_1: \mu \neq 20$$

In other words, until and unless there is an overwhelming evidence against it, he would like to believe that the mean hardness has not changed.

The test statistic is again z, but now he would reject Ho if $\bar{x}$ is too far above 20 as well as if it is too far below 20.

The significance level, $\alpha$ is 0.05 and as shown in Figure VI below, this implies that the total probability of rejecting Ho is 0.05. The critical region now exists under both the tails of the distribution of the test statistic and we would treat both of them to be equal. Therefore, each of the shaded area is 0.025 and one half of the acceptance region has an area 0.475, which corresponds to a z value of 1.96·in normal tables.

Figure VI: Two-talled test of Hypothesise 0.05 Significance level



The lower limit of the acceptance region $= \mu - 1.96\sigma_{\bar{x}}$, when $\mu = 20$

$$= 20 - 1.96\frac{\sigma}{\sqrt{n}}$$

$$= 20 - 1.96 \times \frac{2.5}{\sqrt{49}}$$

= 20 - 0.7

= 19.3

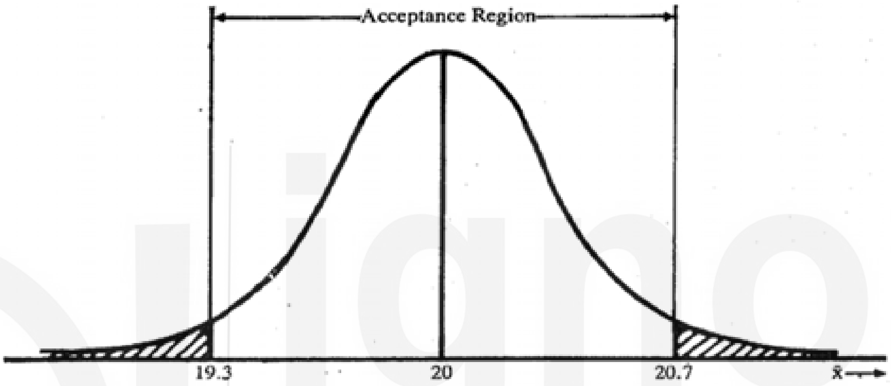
Similarly, the upper limit of the acceptance region $= \mu + 1.96\sigma_{\bar{x}}$

= 20 + 0.7

=20.7

In Figure VII below we have shown the acceptance and the rejection regions. As the observed value of viz. 19.5 falls in the acceptance region, we conclude that the sample evidence is not strong enough for us to reject Ho.

**Figure VII: Acceptance Region for the two-tailed Test**



### When Population Variance is Unknown

We have so far been assuming that the population variance was known and so we could easily calculate the standard error of the mean. However, in many cases that population variance is not known and we still want to draw conclusions about the population mean.

**Sample Size is Large:** When the population standard deviation is not known, we have to estimate it from the sample and as we have discussed in the previous unit we use the sample standard deviation s to estimate the population standard deviation σ. Further, if the sample size is large (n > 30), then the standard error of the mean can be calculated as

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

and so the testing of hypothesis can proceed exactly as in the previous section. It is to be noted that if the population size (N) is small so that the sampling ratio (n/N) is larger than 0.05, then the finite population multiplier also needs to be used for calculating α i.e. such a case

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

**Sample Size is Small:** When the sample size is small (n≤30) and the population standard deviation is unknown, the standard error of mean ($\sigma_{\bar{x}}$) cannot be found directly. However, as we have seen in the previous unit, if the population distribution is normal, the sample standard deviation (s) can be used to calculate the value of a related random variable

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

which has known distribution-viz. the Student's distribution with n – 1 degrees of freedom. Therefore, if the sample standard deviation (s) is known-and this can always be calculated from the sample observations-then the critical region can again be defined in terms of the test statistic sample mean (x). We propose to show how this can be done through an example.

Let us go back to the decision problem faced by the purchase manager as narrated in section 11.4 above-with the only difference that the population standard deviation σ is unknown The purchase manager picks up a sample of size 15 and finds that the sample mean $\bar{x}$ to be 19.5 and the sample standard deviation s as 2.6 , If he uses a significance level of 0.05 as before, can he conclude that the mean hardness of castings from this supplier has changed from 20?

Our null and the alternative hypotheses would remain unchanged, viz.

$$H_0: \mu = 20$$

$$H_1: \mu \neq 20$$

The test statistic is again the sample mean z.

The Sample size is n = 15

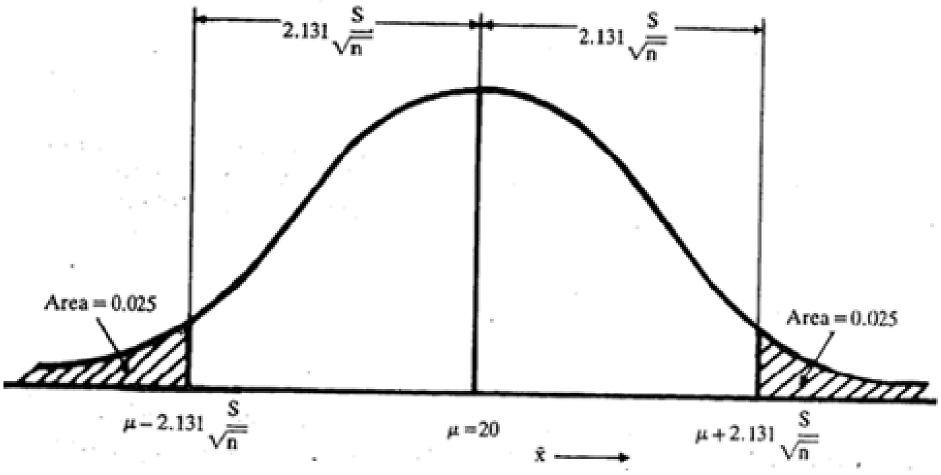
and the observed value of z is 19.5 and that of is 2.6. This is again at two-tailed test and the null hypothesis can be rejected only if the observed value $\bar{x}$ of is too far away From 20-i.e. when $|\bar{x} - 20| \geq c$ where c is a number the value of which depends on the significance level.

The distribution of z is not known directly, but the distribution of a related variable $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is known, when Ho is true-i.e. when $\mu$ = 20. We know that $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ will have a t distribution with (n -1) degrees of freedom and since n = 15, by referring to the t tables, we can see that for a t variable with 14 degrees of freedom,

$$\Pr\,[-2.131 \leqslant t_{14} \leqslant 2.131] = 0.95$$

The symbol $t_{14}$ above represents a t variable 14 degrees of freedom and Figure VIII below shows the critical region for this test. We want that the probability of rejecting Ho when Ho is true-i.e. when $\mu$ = 20, to be 0.05 and this rejection region is under both the tails of the distribution of and so the area under each tail is 0.025 as shown in Figure VIII.

**Figure VIII: Two tailed test of mean for small sample size**


$2.131 \frac{S}{\sqrt{n}}$
$2.131 \frac{S}{\sqrt{n}}$
Area = 0.025
Area = 0.025
$\mu - 2.131 \frac{S}{\sqrt{n}}$
$\mu = 20$
$\bar{x}$ ⟶
$\mu + 2.131 \frac{S}{\sqrt{n}}$

From the t tables, we find that

$$\Pr\left[t_{14} \geqslant 2.131\right] = 0.025$$

But $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ is known to have a t distribution with 14 degrees of freedom.

$$\therefore \Pr\left[\frac{\bar{x}-\mu}{s/\sqrt{n}} \geqslant 2.131\right] = 0.025$$

$$\therefore \Pr\left[\bar{x} \geqslant \mu + 2.131 \frac{s}{\sqrt{n}}\right] = 0.025$$

Finally, if $H_0$ is true, then $\mu = 20$ and so

$$\Pr\left[\bar{x} \geqslant 20 + 2.131 \frac{s}{\sqrt{n}}\right] = 0.025$$

We can, therefore, calculate the limits of the acceptance region as follows:

the upper limit of the acceptance region $= 20 + 2.131 \frac{s}{\sqrt{n}}$

$$= 20 + 2.131 \times \frac{2.6}{\sqrt{15}}$$

= 21.43

and the lower limit of the acceptance region $= 20 - 2.131 \frac{s}{\sqrt{n}}$

$$= 20 - 2 \times \frac{2.6}{\sqrt{15}}$$

= 8.57

But the observed value of x is 19.5, which falls in the acceptance region and so we conclude that the sample evidence is not strong enough for us to reject Ho at a significance level of 0.05.

It is to be noted that we have used a two-tailed test here because that is how our hypotheses were set up. The procedure for a one-tailed test using t distribution is conceptually the same as a one-tailed test using the normal distribution that we have seen earlier in section 11.4 above. Make sure that you are reading the t table correctly because in some t tables the t values for the area under both tails is tabulated whereas in others the t values for the area under one tail only is tabulated.

## 11.5 TESTING OF POPULATION. PROPORTION

We shall now discuss how tests concerning population proportions can be conducted. At this stage, we would request you to review the previous unit where we discussed the determination of confidence interval for the population proportion. In particular, recollect that the sampling distribution of the proportion is actually a binomial distribution, which can be approximated by a normal distribution with the same mean and the same variance if n is sufficiently large so that both np and n(1-p) are at least as large as 5.

A personnel manager wants to know if the competence and the performance of its supervisory staff has changed. He knows from past surveys that 30% of the supervisory staff used to be rated in the "super" category. A sample of 50 supervisory staff have recently been rated and only 12 of them appear in the "super" category. What should the personnel manager conclude at a 5% significance level?

In the absence of an overwhelming evidence against it, the personnel manager is likely to believe that the proportion of supervisory staff in the "super" category has not change. If p is the proportion of supervisory staff in the "super" category in the population, or null and the alterative hypotheses are:

$$\mathrm{H_0: p = 0.3}$$

$$H_1: p \neq 0.3$$

The test statistic is the sample proportion p. If the sample size is large enough [so that both np and n(1- p) are at least as large as 5], then

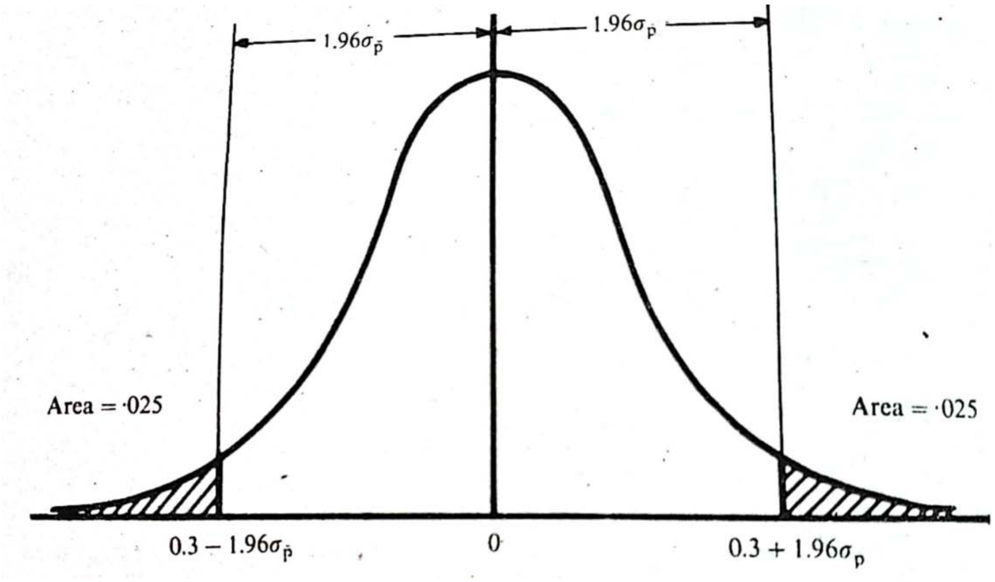
$$\bar{\mathrm{p}} \sim \mathrm{N}\left(\mathrm{p}, \frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}\right)$$

When $\mathrm{H_0}$ is true—i.e. when p = 0.3

$$\bar{\mathrm{p}} \sim N(0.3, 0.0042)$$

In other words, when Ho is true, the sample proportion p approximately follows a normal distribution with mean 0.3 and variance 0.0042.

**Figure IX: A two-tailed test of proportion**



If we represent the standard deviation of the sample proportion $\bar{p}$ as $\sigma_{\bar{p}}$ then, if Ho is true

$$= \sqrt{\frac{p(1-p)}{n}}$$

$$= \sqrt{\frac{0.3 \times 0.7}{50}}$$

= 0.065

From our null and alternative hypotheses, we can easily see that we have a two-tailed test where the null hypotheses will be rejected if the sample proportion p is either too much below or too much above 0.3. We have shown the rejection region in Figure IX above and from normal tables we find that when the area to the right is 0.025, the z value is 1.96. We can, therefore, define the appropriate acceptance region as follows:

the upper limit of

acceptance region $= \mathrm{p} + 1.96\sigma_{\bar{p}}$, when $p = 0.3$

$= 0.3 + 1.96 \times 0.065$

$= 0.43$

and the lower limit of acceptance region $= \mathrm{p} - 1.96\sigma_{\bar{p}}$,

$= 0.3 - 1.96 \times 0.065$

$= 0.17$

In the sample, only 12 out of 50 supervisors belong to the "super" category. So, the observed value of p is

$$\bar{\mathrm{p}} = \frac{12}{50} = 0.024$$

As this value falls in the acceptance region, we conclude that the sample evidence is not strong enough for us to reject Ho and so we accept Ho that the proportion of "super" supervisors has not changed from 0.3.

It is not difficult to see that even with proportions, one can use either a one-tailed test or a two-tailed test (as used above) depending upon how the null and the alternative hypotheses have been set up. The concept and the approach is exactly the same as we have discussed in previous sections and so we are not repeating it here.

**Activity D**

Diagram the acceptance and the rejection regions in each of the following situation where the significance level of the test is 10% and the alternative hypothesis is

1) $H_1: \mu \neq 90$ Drawing 1

2) $H_1: \mu > 90$,, 2

3) $H_1: \mu < 90$,, 3

**Activity E**

| | | | | |
|---|---|---|---|---|
| 1) $H_0: \mu = 90$ $H_1: \mu \neq 90$ | | $\bar{x} = 92.1$ | $\sigma = 8$ | n = 36 |
| $\bar{x} \sim N\left(90, \frac{64}{36}\right)$ when $\mu = 90$ | | | | |
| 2) $H_0: \mu = 90$ | $H_1: \mu \neq 90$ | $\bar{x} = 92.1$ | S = 8 | n =36 |
| 3) $H_0: \mu = 90$ | $H_1: \mu \neq 90$ | $\bar{x} = 92.1$ | S = 8 | n=20 |
| 4) $H_0: p = 0.3$ | $H_1: p \neq 0.3$ | $\bar{p} = 0.38$ | | n = 50 |

In each of the following cases, specify which probability distribution you would use to conduct the test:

## 11.6 TESTING FOR DIFFERENCE BETWEEN MEANS

Many a time the decision maker is interested in knowing whether two related populations are different from each other in respect of any parameter of the population. For example, a marketing manager may be interested in knowing whether the mean sales from a retail shop is affected by a display at the point of purchase. A personnel manager may like to know whether the job performance of a category of employees is affected by a particular training programme. In these cases, the decision maker is not interested in concluding anything about the parameter value in either of the populations, but only whether the difference is significant or not. We shall study testing for difference between two means in this section. In the following section, we shall take a look at testing for the difference between proportions.

**Independent Samples**

We first discuss the case where we want to arrive at some conclusion about the difference between two population means and we draw one sample from each of the populations, independent of the other. So, we have two independent samples and we want to test the difference between the two population means based on the evidence produced by the two samples.

**Sampling Distribution of the Difference between Sample Means: Let us assume that**

the mean and variance of the first population are $\mu_1$, and $\sigma_1^2$ respectively, and similarly, let $\mu_2$, and $\sigma_2^2$ be the mean and variance of the second population.

Let $\bar{x}_1$ be the sample mean of a sample of size $n_1$ from the first population and $\bar{x}_2$ the sample of a sample of size $n_2$ from the second population.

From our earlier discussion on the sampling distribution of the mean, we know that

$$E(\bar{x}_1) = \mu_1$$

$$\text{and Var } (\bar{x}_1) = \frac{\sigma_1^2}{n_1}$$

if the first population is not so small as to need the finite population multiplier.

Similarly, $E(\bar{x}_2) = \mu_2$

and Var $(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$

Now, if the samples are independent, the random variables $x_1$ and $x_2$ are also independent and so

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

and Var $(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Finally, if $x_1$ and $x_2$ are normally distributed, then the difference between these two random variables would also be normally distributed. In other words.

$$\text{if } \bar{x}_1 - N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\text{and } \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^{\frac{2}{2}}}{n_2}\right)$$

then, $(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

**Tests When Sample Sizes are Large**: When $n_1$ & $n_2$ are large, we know from the Central Limit Theorem that both $\bar{x}_1$ and $\bar{x}_2$ would be normally distributed. If $\sigma_1$ and $\sigma_2$ are known, then the distribution of $(\bar{x}_1-\bar{x}_2)$ is also known completely and one can directly proceed with tests concerning $(\mu_1-\mu_2)$. On the other hand, even if $\sigma_1$ and $\sigma_2$ are not known, they can be easily estimated by the respective sample standard deviations and one can proceed as if the population standard deviations are known. We shall now demonstrate this procedure by an example.

A marketing manager wants to know if display at point of purchase helps in increasing the sales of his product. Unless there is strong evidence to the contrary, he is likely to believe that such displays do not affect sales. He picks up 70 retail shops where there is no display and finds that the weekly sale in these shops has a mean of Rs. 6000 and a standard deviation of Rs. 1004. Similarly, he picks up a second sample of 36 retail shops with display at point of purchase and finds that the weekly sale in these shops has mean of Rs. 6500 and a standard deviation of Rs. 1200. What should he conclude at a significance level of 5%?

Let us use the subscript 1 to denote the first population (i.e. without display) and subscript 2 for the second population (i.e. with display). The null and the alternative hypotheses follow:

$$\mathrm{H}_0\colon \mu_1 \geqslant \mu_2 \text{ i.e. } \mu_1-\mu_2 \geqslant 0$$

$$\mathrm{H}_1\colon \mu_1 < \mu_2 \text{ i.e. } \mu_1-\mu_2 < 0$$

In the absence of strong evidence to the contrary, he is likely to. accept that display does not increase sales. The test statistic to be used is $(\bar{x}_1-\bar{x}_2)$ and since both $\mathrm{n}_1$ and $\mathrm{n}_2$ are large,

$$(\bar{x}_1-\bar{x}_2)\sim N\left(\mu_1-u_2,\frac{\sigma_1^2}{n_1}+\frac{\alpha_2^2}{n_2}\right)$$

The probability of type I error is the highest when $(\mu_1-\mu_2)$ is at the breakpoint value between $\mathrm{H}_\mathrm{o}$ and $H_1$i.e. when $\mu_1=\mu_2$ and so

$$(\bar{x}_1-\bar{x}_2)\sim N(0,\sigma^2\bar{x}_1-\bar{x}_2)$$

where $\sigma_{\mathrm{x}_1-\bar{x}_2}$ is the standard deviation of $\left(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2\right)$. We can easily calculate $\sigma_{\bar{x}_1-\bar{x}_2}$ by substituting $s_1$ for $\sigma_1$ and $s_2$ for $\sigma_2$ as both $n_1$ and $n_2$ are large.

$$\therefore \sigma_{\bar{x}_1-\bar{x}_2}^2=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}=\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$$

$$=\frac{1004\times 1004}{70}+\frac{1200\times 1200}{30}$$

$$=14400+40000$$

$$=54400$$

$$\therefore \sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{54400}=233.24$$

So, we know that when $(\mu_1-\mu_2)$ is at the breakpoint value between $\mathrm{H}_0$ and $\mathrm{H}_1$, $(\bar{x}_1-\bar{x}_2)$ is normally distributed with a mean 0 and a standard deviation of 233.24

ignou
THE PEOPLE'S
UNIVERSITY

i.e.$(\bar{x}_1 - \bar{x}_2) \sim N(0,54400)$

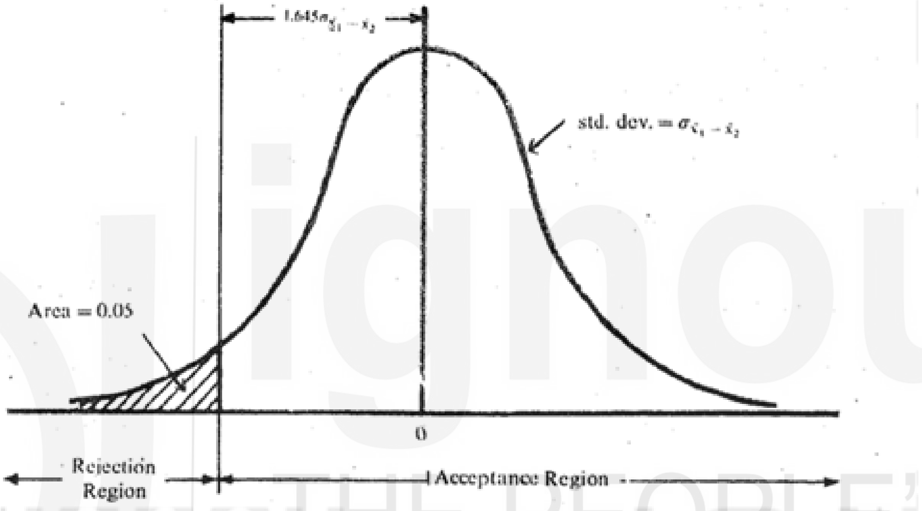
We can reject $H_0$ only when $(\bar{x} \quad - \bar{x}_2)$ is sufficiently negative so that the probability of getting a value as small as the cut-off value is not larger than the significance level 0.05. From normal tables, we find the corresponding z value to be $-1.645$ and so, as shown in Figure X below, the cut-off value of

$$(\bar{x}_1 - \bar{x}_2) = 0 - 1.645\sigma_{\bar{x}_1 - \bar{x}_2}$$

$$= -\,1.645 \times 233.24$$

$$= -\,383.68$$

**Figure X: Testing for difference between means: large independent samples**



The test procedure can, therefore, be summarised as

Reject $H_o$ if $(\bar{x}_1 - \bar{x}_2) \leqslant -383.68$

Our observed value of $X_1$ is 6000 and that of $\overline{X}_2$ is 6500 and so the observed value of $(\bar{x}_1 - \bar{x}_2) = -500$ and so we can reject $H_o$ at 5% significance level and conclude that display at point of purchase does increase sales.

This test turned out to be a one-tailed test, but even when the null and the alternative hypotheses are such that we have a two-tailed test, the approach is similar to the two tailed tests that we have discussed earlier.

**Tests When Sample Sizes are Small:** When the sample sizes $n_1$ and $n_2$ are small, we cannot substitute $s_1$ for $\sigma_1$ and $s_2$ for $\sigma_2$ and proceed as if $\sigma_1$ and $\sigma_2$ are known. We shall develop a procedure for this case here, when we can make the further assumption that $\sigma_1 = \sigma_2 = \sigma$ (say). If $\sigma_1$ and $\sigma_2$ are known to be different, such a situation is beyond the scope of this course.

Having assumed that $\sigma_1 = \sigma_2 = \sigma$, our estimate for a is a pooled standard deviation $s_p$ defined as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

We could have estimated a by $s_1$ or $s_2$ alone but then we would not have used all the information available to us. Using $s_p$ as our estimate of the standard deviation of the two populations, the estimate of the standard deviation of the difference between the two sample means works out to

$$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

And finally, when σ is replaced by $s_p$, the distribution of

$$\frac{\left(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2\right)-(\mu_1-\mu_2)}{\mathrm{s_p}\sqrt{\frac{1}{\mathrm{n}_1}+\frac{1}{\mathrm{n}_2}}}$$

is a t distribution with $(\mathrm{n}_1+\mathrm{n}_2-2)$ degrees of freedom. We can, therefore, develop a test procedure using the t distribution with $(\mathrm{n}_1+\mathrm{n}_2-2)$ degrees of freedom as shown in the example below.

Let us take up the decision problem faced by the marketing manager in this section where he wants to know if display at point of purchase helps in increasing sales. He picks up 12 retail shops with no display and finds that the weekly sale in these shops has a mean of Rs. 6000 and a standard deviation of Rs. 1004. Similarly, he picks up a second sample of 10 retail shops with display at point of purchase and finds that the weekly sale in these shops has a mean of Rs. 6500 and a standard deviation of Rs. 1200. What should he-conclude at a significance level of 5%?

We first state the null and the alternative hypothesis as follows:

$$\mathrm{H}_0: \mu_1 \geqslant \mu_2 \text{ i.e. } \mu_1-\mu_2 \geqslant 0$$

$$\mathrm{H}_1: \mu_1 < \mu_2 \text{ i.e. } \mu_1-\mu_2 < 0$$

where the symbols have the same meaning as in this section above.

The test statistic will again be $\left(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2\right)$ and if the population are normally distribute I then $\left(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2\right)$ will also have a normal distribution with its mean as $(\mu_1-\mu_2)$ and a standard deviation which can be estimated by the pooled standard deviation

$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

We know that $\mathrm{n}_1=12, \mathrm{s}_1=1004$

and $\mathrm{n}_2=10, \mathrm{s}_2=1200$

$$s_p=\sqrt{\frac{(12-1)\times(1004)^2+(10-1)\times(1200)^2}{12+10-2}}$$

$$= \sqrt{\frac{1{,}10{,}88{,}176 + 1{,}29{,}60{,}000}{20}}$$

$$= \sqrt{\frac{2{,}40{,}48{,}176}{20}}$$

$$= \sqrt{1202408.8} = 1096.54$$

so $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ will have a t distribution with

$(n_1 + n_2 - 2)$ degrees of freedom. Since the significance level is 5%, the probability of type I error should not exceed .05 and as shown in Figure XI below, we find from t tables the probability that a t variable with (12 + 10- 2) i.e. 20 degrees of freedom takes α value as small as - 1.725 is .05. The probability of type I error is the highest when $(\mu_1 - \mu_2)$ is at the breakpoint value between Ho and $H_1$-i.e. When $(\mu_1 - \mu_2) = 0$

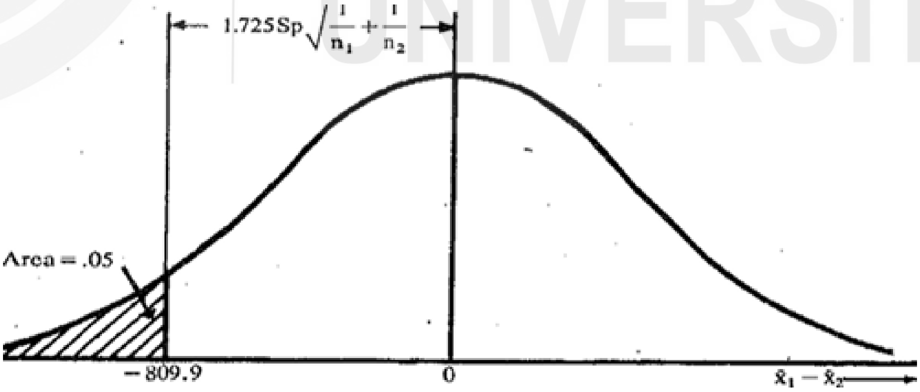
and so the cut-off value of $(\bar{x}_1 - \bar{x}_2)$ would be given by

$$(\mu_1 - \mu_2) - 1.725 \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ when } (\mu_1 - \mu_2) = 0$$

$$= 0 - 1.725 \times 1096.54 \times \sqrt{\frac{1}{12} + \frac{1}{10}}$$

$$= -809.9$$

**Figure XI: One-tailed test of difference between means: small independent samples**



The test procedure can, therefore, be summarised as:

Reject $H_0$ if $(\bar{x}_1 - \bar{x}_2) \leq -809.9$

Our observed value of $x_1$ is 6000 and that of $x_2$ is 6500 and so the observed value of $(\bar{x}_1 - \bar{x}_2) = -500$ and as this belongs to the acceptance region, we conclude that the Evidence is not strong enough for us to reject $H_o$ That is, we accept the null hypothesis that display at point of purchase does not increase sales.

### Dependent Samples

We have so far discussed the case when the two samples picked up from the populations were independent-but we can also design our test in such a way that the samples are dependent. For example, if we want to know whether a training programme helps in improving the job performance of a category of employees, we can evaluate the job performance of a sample of employees before they have undergone the training programme. We can evaluate the performance of the employees again-after they have undergone the training programme. We would, therefore, have two performance evaluations for each employee in our sample-one before and the other after the training programme and so the two samples are dependent on each other. For each employee the difference in the performance evaluations is caused by the training programme and many other random factors which have a very insignificant effect on the job performance. Therefore, the difference in the performance evaluations can be treated as a random variable having a distribution of its own.

In general, using dependent samples is better than using independent samples because the effect of all other major factors is eliminated and the difference can be attributed only to the "treatment" that we are studying. Such a design may not always be possible but whenever we can design a test based on dependent samples, we are relatively more confident that we have isolated the effect of the "treatment" and that the two samples are identical but for this difference in "treatment".

We shall again consider the decision problem faced by the marketing manager in 11.6 above regarding whether display at point of purchase helps in increasing sales. He picks up a random sample of 11 retail shops and notes down the weekly sales in each of these shops. Next, he introduces display at point of purchase at each of these shops and again observes the weekly sales in them, as given in Table 2 below. If he is using α significance level of 5%, what should he conclude?

Using the same symbols as earlier, we introduce one more random variable, d, defined as

$$d = x_1 - x_2$$

i.e. d is the difference in sales in a retail shop between before and after the display. If the expected value of d is represented by $\mu_d$, then

$$\mu_d = \mu_1 - \mu_2$$

Let us write our null and the alternative hypotheses as before:

$$H_0 : \mu_1 - \mu_2 \geqslant 0 \text{ i.e. } \mu_d \geqslant 0$$

$$H_1 \quad : \mu_1 - \mu_2 < 0 \quad \text{ i.e. } \mu_d < 0$$

As you can see this is a test concerning the population mean when we have a sample of d values. We use the sample mean d as the test statistic and because the sample size is small (n=11), we shall use a t test.

**Table 2: Weekly Sales in a Sample of 11 Retail Shops**

| Shop No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before Display | 4500 | 5275 | 7235 | 6844 | 5991 | 6672 | 4943 | 7615 | 6148 | 5623 | 5154 |
| After Display | 4834 | 5010 | 7562 | 6957 | 6401 | 6423 | 5334 | 8004 | 6729 | 6277 | 5769 |
| Difference (d) | -334 | 265 | -327 | -113 | -410 | 249 | -391 | -389 | -581 | -654 | -615 |

From the sample, we find that for n =11 the sample mean d = - 300 and the sample standard deviation, $s_d$ = 314.53.

If we assume that the d values are normally distributed, then the cut-off value can be easily obtained from the t tables with (11, 0.05) degrees of freedom, as shown in Figure XII below.

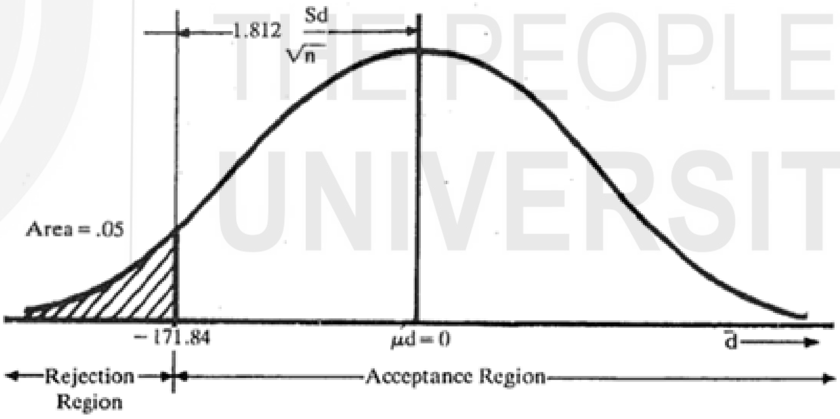
**The cut-off value of**

$$a = \mu_{\mathrm{d}} - 2.812\frac{\mathrm{S_d}}{\sqrt{11}}, \text{ when } \mu_{\mathrm{d}} = 0$$

$$= 0 - 1.812 \times \frac{314.53}{\sqrt{11}}$$

$$= -171.84$$

**Figure XII: One-tailed test of difference between means: small dependent samples**



As our observed value of d' is - 300, it is very much in the rejection region and so we can conclude that display at point of purchase does increase sales. We can also see that if the sample size is large, we can use the z test in place of the t test. Also, that both one- and two-tailed tests can be performed depending upon the hypotheses that are set up.

## 11.7 TESTING FOR DIFFERENCE BETWEEN PROPORTIONS

A marketing manager wants to know if there is any difference in the proportion of consumers who like the taste of his product. He finds that 40 out of a sample of 85 consumers respond that they like the taste of his product. Similarly, 35 out of a second sample of 65 consumers respond that they like the taste of the product-when they are administered a product of the next competing brand. Based on these observations, what should the marketing manager conclude at a 5% significance level?

Let us first state the null and the alternative hypotheses:

$$\mathrm{H_O}: \bar{\mathrm{p}}_1 = \bar{\mathrm{p}}_2$$

$$\mathrm{H_1}: \bar{\mathrm{p}}_1 \neq \bar{\mathrm{p}}_2$$

where $\bar{p}_1$ refers to the proportion of consumers who like the product of the marketing manager and $\bar{p}_2$ the proportion of consumers who like the product of the next competing brand. The test statistic will be $\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2$i.e. the difference in the two sample proportions. Since the sample sizes $n_1$ and $n_2$ are large enough

$$\bar{\mathrm{p}}_1 \sim \mathrm{N}\left(\bar{\mathrm{p}}, \frac{\bar{\mathrm{p}}_1(1-\bar{\mathrm{p}}_1)}{\mathrm{n}_1}\right)$$

and $\bar{\mathrm{p}}_2 \sim \mathrm{N}\left(\mathrm{p}, \frac{\bar{\mathrm{p}}_2(1-\bar{\mathrm{p}}_2)}{\mathrm{n}_2}\right)$

And since the two samples are independent.

$$(\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2) \sim \mathrm{N}\left(\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2, \frac{\bar{\mathrm{p}}_1(1-\bar{\mathrm{p}}_1)}{\mathrm{n}_1} + \underline{\bar{\mathrm{p}}}_2 \frac{(1-\bar{\mathrm{p}}_2)}{n_2}\right)$$

The significance level being 0.05, we would like the probability of rejecting Ho when Ho is true to not exceed 0.05 and so, as shown in Figure XIII below

to upper limit of the acceptance region

$$= (\bar{p}_1 - \bar{p}_2) + 1.96\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}, \text{ when } \bar{p}_1 = \bar{p}_2$$

$$= 0 + \frac{1.96\sqrt{\bar{\mathrm{p}}_1(1-\bar{\mathrm{p}}_1)}}{\mathrm{n}_1} + \frac{\bar{\mathrm{p}}_2(1-\bar{\mathrm{p}}_2)}{\mathrm{n}_2}$$

We shall substitute $p_1$ and $P_2$ by their estimates $\bar{p}_1$ and $\bar{p}_2$. However, when $\bar{\mathrm{p}}_1 = \bar{\mathrm{p}}_2 = \mathrm{p}$ (say), it would be even better to have a pooled estimate of p, say $\bar{p}$ from both the samples put together.

$$\therefore \bar{\mathrm{p}} = \frac{\mathrm{n}_1\bar{\mathrm{p}}_1 + \mathrm{n}_2\bar{\mathrm{p}}_2}{n_1 + \mathrm{n}_2}$$

$$= \frac{40+35}{85+65}$$

$$= \frac{75}{150} = 0.50$$

and so the upper limit ,

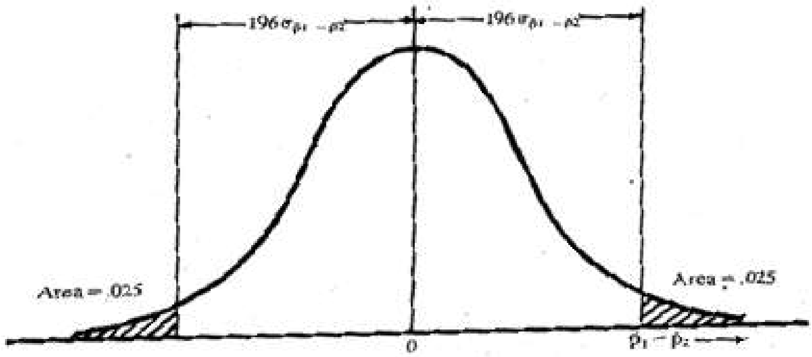
of the acceptance region $= 0 + 1.96\sqrt{\bar{\mathrm{p}}(1-\bar{\mathrm{p}})\left(\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}\right)}$

ignou
THE PEOPLE'S
UNIVERSITY

$$= 1.96\sqrt{0.5 \times 0.5 \times \left(\frac{1}{85} + \frac{1}{65}\right)}$$

$$= 1.96 \times 0.082 = 0.16$$

**Figure XIII: Test for difference of population proportions**


1.96 $\sigma_{\bar{p}_1-\bar{p}_2}$
1.96 $\sigma_{\bar{p}_1-\bar{p}_2}$
Area = .025
Area = .025
0
$\bar{p}_1 - \bar{p}_2$

Similarly, the lower limit

of the acceptance region $= 0 - 1.96 \times \sigma_{\bar{p}_1 - \bar{p}_2}$

$= 1.96 \times 0.082$

$= -0.16$

The observed value of $(\bar{p}_1 - \bar{p}_2)$ is

$$(\bar{p}_1 - \bar{p}_2) = \frac{40}{85} - \frac{35}{65}$$

$$= 0.47 - 0.54$$

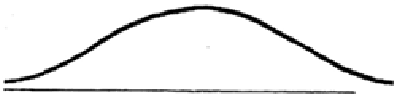
$$= -0.07$$

As the observed value of $(\bar{p}_1 - \bar{p}_2)$ falls in the acceptance region, we conclude that the sample evidence is not strong enough for us to reject $H_o$. Similar tests can also be conducted when the null and the alternative hypotheses are so set up that one-tailed tests are required.

### Activity F

Diagram the acceptance and the rejection regions in each of the following situations when the significance level of the test is 10% and the alternative hypotheses are

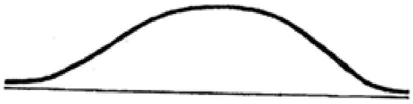
a) $\mathbf{H}_1: \mu_1 - \mu_2 \neq 0$
(independent samples)



b) $\mathbf{H}_1: \boldsymbol{\mu_1} - \boldsymbol{\mu}_2 > \mathbf{0}$
(independent samples)

c) **$\mathbf{H}_1$: $\boldsymbol{\mu}_1 - \mu_2 \nless 0$**
(independent samples)

d) **$\mathbf{H}_1$**: $\mu_1 - \mu_2 \neq 0$
(dependent samples)



e) **$\mathbf{H}_1: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 > 0$**
(dependent samples)

f) **$\mathbf{H}_1: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 < 0$**
(dependent samples)

**Activity G**

In each of the following cases, specify which probability distribution you would use to conduct the test:

a) $H_0: \mu_1 - \mu_2 = 0$ $\bar{x}_1 = 32.1$ $\sigma_1 = 7.5$ $n_1 = 13$
$H_1: \mu_1 - \mu_2 \neq 0$ $\bar{x}_2 = 36.2$ $\sigma_2 = 6.7$ $n_2 = 11$
(independent samples)

b) $H_0: \mu_1 - \mu_2 = 0$ $\bar{x}_1 = 32.1$ $s_1 = 7.5$ $n_1 = 13$
$H_1: \mu_1 - \mu_2 \neq 0$ $\bar{x}_2 = 36.2$ $s_2 = 5.7$ $n_2 = 11$
(dependent samples) $d = -4.1$ $S_d = 3.7$

c) $H_0: \mu_1 - \mu_2 = 0$ $\bar{x}_1 = 32.1$ $s_2 = 7.5$ $n_1 = 12$
$H_1: \mu_1 - \mu_2 \neq 0$ $\bar{x}_2 = 36.2$ $s_2 = 6.7$ $n_2 = 12$
(dependent samples) $d = -4.1$ $s_d = 3.7$

d) $H_0: p_1 - p_2 = 0$ $\bar{p}_1 = 0.35$ $n_1 = 20$
$H_1: p_1 - p_2 \neq 0$ $\bar{p}_2 = 0.40$ $n_2 = 30$

## 11.8 SUMMARY

In this unit we have seen how tests concerning statistical hypotheses can be designed and used. A statistical hypothesis is a statement about a population parameter or about a population distribution. As these tests are conducted on the basis of evidence thrown up by a sample, errors cannot he totally eliminated. All tests are designed to answer the question- "Is the sample evidence strong enough to reject the null hypothesis?". The null and the alternative hypotheses are set up such that one of them, and only one of them, is always True. In the absence of a strong evidence to the contrary, the decision maker would be willing to accept the null hypothesis.

Of the two errors that are possible in any testing of hypothesis, type I error- viz. the error in wrongly rejecting the null hypothesis-is considered to be more serious than the other one and so is subject to explicit control. All tests

ignou
THE PEOPLE'S
UNIVERSITY

are performed at a significance level which defines the highest probability of type I error.

All tests of hypotheses are conducted in two phases-in the first phase a test is designed where we decide as to when can the null hypothesis be rejected-and in the second phase the designed test is used to draw the conclusion.

We then looked at some specific test. We found that while testing population means, the test can be based on the normal distribution if the population variance was known or if the sample size was large. On the other hand, if the sample size was small, we had to design a test based on the t distribution. Population proportions could also be tested on the basis of normal distribution.

We then developed tests for testing the difference between two population means- both for independent and for dependent samples. When the samples were independent and the sample sizes were small, we developed a t test based on the pooled estimate of the standard deviation of the two populations, under the assumption that they were equal. Similarly, we also developed a .procedure for testing the difference between two population proportions.

## 11.9 SELF-ASSESSMENT EXERCISES

1) A personnel manager has received complaints that the stenographers in the company have become slower and do not have the requisite speeds in stenography. The Company expects the stenographers to have a minimum speed of 90 words per minute. The personnel manager decides to conduct a stenography test on a random sample of 15 stenographers. However, he is clear in his mind that unless the sample evidence is strongly against it, he would accept that the mean speed is at least 90 w.p.m. After the test, it is found that the mean speed of the 15 stenographers tested is 86.2 w.p.m. What should the personnel manager conclude at a significance level of 5%, if it is known that the standard deviation of the speed of all stenographers is 10 w.p.m.

2) The marketing manager of a firm has decided to launch a new ready-to-eat snack. There are two minor variations of the product which have been developed. Both of these are basically similar but a bit different in their colour, flavour and crispness. Also, both of these are highly perishable and have a shelf life of about 48 hours.

   The marketing manager decides to conduct a field trial of both the product variants to find out if one is liked better by. the people as compared to the other. He selects 20 shops which are similar in respect of their sizes, locations, clientele; etc. He introduces the first variant of the product (say Pr) in 12 of these shops and similarly, he introduces the second variant (say $P_2$) in the other 8. Complete records are kept of the movement of these products for 15 days. The total sales of $P_1$ and $P_2$ in these shops in a period of 15 days is found to be as follows:

(Sale in kg)

For $P_1$ : 25 23 19 18 23 20 21 19 22 24 25

For $P_2$ : 22 19 . 23 22 17 20 22 19

Both $P_1$ and $P_2$ are priced equally. The marketing manager now wants to conclude whether there is any significant difference between $P_1$ and $P_2$. Using a significance level of 1%, what can he conclude?

3) The situation is the same as in 2 above. However, suppose that instead of selecting 20 shops, the marketing manager selects only 10 shops and he introduces both the products in all the 10 shops. At the end of 15 days, he finds that the total sales in each of these 10 shops has been as follows:

(Sale : in kg)

| Shop | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Product $P_1$ | 14 | 17 | 12 | 9 | 13 | 15 | 13 | 13 | 10 | 9 |
| Product $P_2$ | 12 | 12 | 12 | 11 | 16 | 12 | 16 | 17 | 10 | 11 |

What should his conclusion be?

4) The currently used manufacturing process is known to produce 5% defectives which is considered to be too high by the management. An alternative process had been suggested and the management wants to get a sample of some components produced by the alternative process, which is operational at another location: What are the null and the alternative hypotheses relevant for this situation? Please discuss why.

For each of the following statements, choose the most appropriate response from among the listed ones:

5) The significance level is probability based on the assumption that

a) Ho is True

b) Ho is False

c) the population mean is known

d) the population variance is known

6) An observed sample for a test of hypothesis yields a P value of 0.075. For this situation, at $\alpha = 0.05$

a) we reject Ho

b) we accept Ho

c) acceptance of Ho depends on whether we have, a one-or two tailed test.

d) we can neither accept nor reject Ho.

7) Testing of hypothesis has some similarities with legal proceedings where, guilt needs to be proven "beyond a reasonable doubt". If innocence were considered to be the null hypothesis, "reasonable doubt" would be quantified by

a) $1 - \alpha$

b) P value

c) R

d) $\alpha$

8) The major purpose of a test of hypothesis is to

   a) make a decision about the sample, using the statistic

   b) make a decision about the observed statistic

   c) make a decision about the population, using the statistic

   d) none of the above.

## 11.10 FURTHER READINGS

Bruce Bomerman, Business Statistics for Practices, McGraw Hill.

Gravetter, F.J. and L.B. Wallnau. *Statistics for the Behavioural Sciences*, West Publishing Co.: St. Paul.

Levin, R.I., *Statistics for Management,* Prentice-Hall of India: New Delhi.

Mason, R.D., *Statistical Techniques in Business and Economics*, Richard D. Irwin, Inc: Homewood.

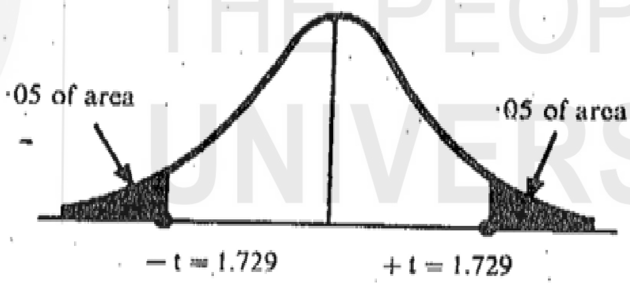
Mendenhall, W. Scheaffer, R.L. and D.D. Wacl erly. *Mathematical Statistics with Applications,* Duxbury Press: Boston.

Plane, D.R. and E.B. Oppermann. *Business and Economic Statistics*, Business Publications Inc.: Plano.

ignou
THE PEOPLE'S
UNIVERSITY

### t DISTRIBUTION

**Areas in Both Tails Combined for Student's t Distribution**



**EXAMPLE:** To find the value oft which corresponds to an area of .10 in both tails of the distribution, combined, when there are 19 degrees of freedom, look under the .10 column, and proceed down to the 19 degrees of freedom now; the appropriate t value there is 1,729.

| Degrees of freedom | .10 | .05 | .02 | .01 |
|---|---|---|---|---|
| 1 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.812 | 2.228 | 2.764 | 3.169 |

| | | | | |
|---|---|---|---|---|
| 11 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 1.658 | 1.980 | 2.358 | 2.617 |
| Normal Distribution | 1.645 | 1.960 | 2.326 | 2.576 |

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 12 CHI-SQUARE TESTS

**Objectives**

By the time you have successfully completed this unit, you should be able to:

- appreciate the role of the chi-square distribution in testing of hypotheses
- design and conduct tests concerning the variance of a normal population
- perform tests regarding equality of variances from two normal populations
- have an intuitive understanding of the concept of the chi-square statistic
- use the chi-square statistic in developing and conducting tests of goodness of fit and
- tests concerning independence of categorised data.

**Structure**

12.1 Introduction
12.2 Testing of Population Variance
12.3 Testing of Equality of Two Population Variances
12.4 Testing the Goodness of Fit
12.5 Testing Independence of Categorised Data
12.6 Summary
12.7 Self-assessment Exercises
12.8 Further Readings

## 12.1 INTRODUCTION

In the previous unit you have studied the meaning of testing of hypothesis and also how some of these tests concerning the means and the proportions of one or two populations could be designed and conducted. But in real life, one is not always concerned with the mean and the proportion alone-nor is one always interested in only one or two populations. A marketing manager may want to test if there is any significant difference in the proportion of high income households where his brand of soap is preferred in North, South, East, West and Central India. In such a situation, the marketing manager is interested in testing the equality of proportions among five different populations: Similarly, a quality control manager may be interested in testing the variability of a manufacturing process after some major modifications were carried out on the machinery vis-a-vis the variability before such modifications. The methods that we are going to introduce and discuss in this unit will help us in the kind of situations mentioned above as well as in many other types of situations. Earlier (section 11.6 in the previous unit), while testing the equality of means of two populations based on small independent samples, we had assumed that both the populations had the same variance and, if at all, their means alone were different. If required, the equality of variances could be tested by using methods to be discussed in this unit.

In many of our earlier tests, we had assumed that the population distribution was normal. It should be possible for us to test if the population distribution is really normal, based on the evidence provided by a sample. Similarly, in another situation it should be possible for us to test whether the population distribution is Poisson, Exponential or any other known distribution.

Finally, the procedures to be discussed in this unit also allow us to test if two variables are independent when the data is only categorised we may, for instance, like to test whether consumer performance for a brand and income level are independent-i.e. the variables e.g. the sex of respondents, have been measured only grouping respondents in categories.

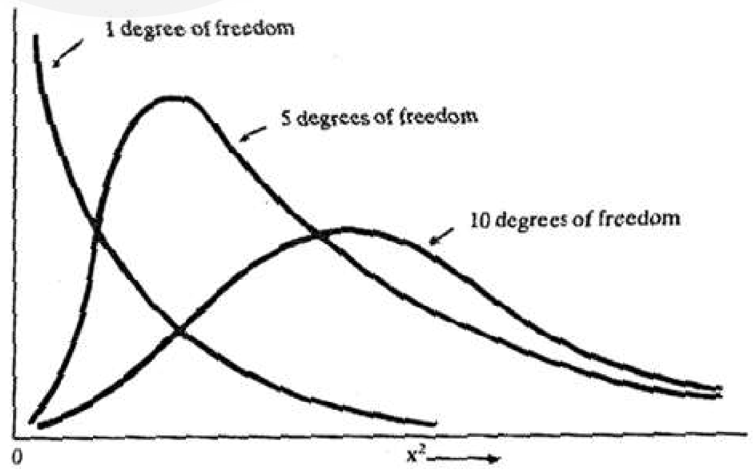
The common thread running through all the diverse situations mentioned above is the chi-square distribution first introduced to you in section 14.4 of unit 14. We start with a recapitulation of the chi-square distribution below before we start with the statistical tests.

### The Chi-Square Distribution--A Recapitulation

A chi-square distribution is known by its only parameter viz. the degrees of freedom. Figure I below shows the probability density function of some chi-square distributions. The left and the right tails of chi-square distributions with different degrees of freedom are extensively tabulated.

If x is a random variable having a standard normal distribution, then $x^2$ will have a chi-square distribution with one degree of freedom. If $Y_1$ and $Y_2$ are independent random variables having chi-square distributions with $v_1$ and $v_2$ degrees of freedom respectively, then $(Y_1 + Y_2)$ will have a chi-square distribution with $v_1 + v_2$ degrees of freedom.

**Figure I: Chi-square distributions with different degrees of freedom**



As shown in Figure I above, if $x^2$ is a random variable having a chi-square distribution with v degrees of freedom, then $x^2$ can assume only non-negative values. Also, the expectation and the variance of $x^2$ is known in terms of its degrees of freedom as below:

$$E[x^2] = v$$

and var $[x^2] = 2v$

Finally, if $x_1, x_2 \ldots, x_n$ are n random variables from a normal population with mean $\mu$ and variance $\sigma^2$ and if the sample mean x and the sample variance $s^2$ are defined as

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Then, $\frac{(n-1)s^2}{\sigma^2}$ will have a chi-square distribution with (n -1) degrees of freedom. Although the distribution of sample variance ($s^2$) of a random sample from a normal population is not known explicitly, the distribution of a related random variable viz $\frac{(n-1)s^2}{\sigma^2}$ is known and is used.

## 12.2 TESTING OF POPULATION VARIANCE

Many times, we are interested in knowing if the variance of a population is different from or has changed from a known value. As we shall see below, such tests can be easily conducted if the population distribution is known to be or can be assumed to be normal. We shall develop and use the test procedure under different null and alternative hypotheses.

**One-Tailed Test**

The specifications for the surface hardness of a composite metal sheet require that the surface hardness be uniform to the extent that the standard deviation should not exceed 0.50. A small random sample of sheets is selected from each shipment and the shipment is rejected if the sample variance is found to be too large. However, a shipment can be rejected only when there is an overwhelming evidence against it. The sample variance from a sample of nine sheets worked out to 0.32. Should this shipment be rejected at a significance level of 5%?

It is clear that in absence of a strong evidence against it, the shipment should be accepted and so the null and the alternative hypotheses should be:

$$H_0: \sigma^2 \leqslant 0.25$$

$$H_1: \sigma^2 > 0.25$$

The highest acceptable value of v is 0.50 and so the highest acceptable value of σ2 is 0.25. If the true variance of the population (shipment) is above 0.25, then the alternative hypothesis is true. However, in the absence of a strong evidence against it, the null hypothesis cannot be rejected and so the shipment will be accepted.

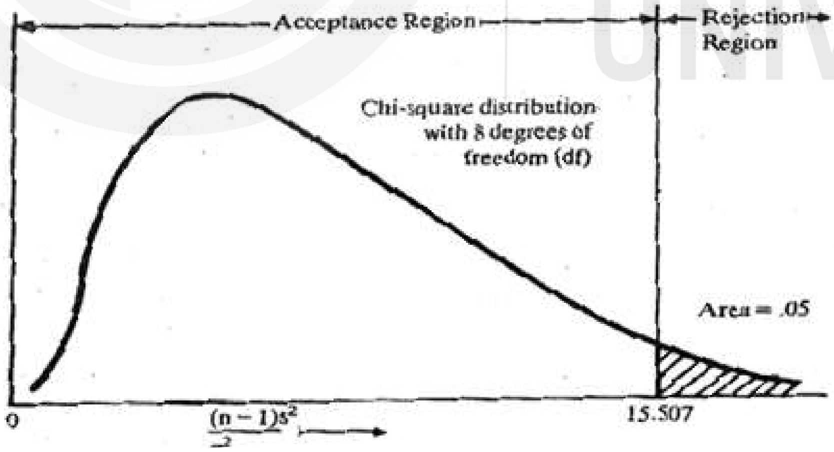
We assume that the surface hardness of these composite metal sheets is distributed normally. The test statistic that we shall use would ideally be the sample variance, but Since the distribution of $s^2$ is not known directly. We shall use $\frac{(n-1)s^2}{\sigma^2}$ as the test statistic which is known to have a chi-square distribution with (n -1) degrees of freedom.

We shall reject the null hypothesis only when the observed value of $s^2$ is much larger than $\sigma^2$. Suppose we reject the null hypothesis if $s^2 > c$, where c is a number much larger than $\sigma^2$, then the probability of type I error should not exceed .05-the given significance level of the test. As before, the probability of type I error is the highest when $v^2$ is at the breakpoint value between $H_0$ and $H_1$ i.e. when $\sigma^2 = 0.25$ Therefore, $\Pr\,[s^2 > c] = 0.05$, when $\sigma^2 = 0.25$

or, $\Pr\left\{\frac{(n-1)s^2}{r^2} > \frac{(n-1)c}{0.25}\right\} = 0.05$

Since $\frac{(n-1)s^2}{\sigma^2}$ is known to have a chi-square distribution with (n -1) –degrees of freedom, we can refer to the tables for the chi-square distribution where the left tail and the right tail are tabulated separately for different areas tinder the tail. As shown in Figure II below, the probability that a $\chi^2$ variable with (9 -1) = 8 degrees of freedom will assume values above 15.507 is 0.05. So,. if the (observed) value of $\chi^2$, i.e, the value of $\chi^2$ calculated, from the observed value of $s^2$ when $\sigma^2 = 0.25$, is greater than 15.507, then only can we reject the null hypothesis at a significance level of .05.

**Figure II: Rejection region for a one-tailed Test of Variance**



The observed value of $s^2$ has been 0.32. So, the observed value of $x^2$ has been

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(9-1) \times 0.32}{0.25}$$

$$= 10.24$$

As this is smaller than the cut-off value of 15.507, we conclude that we do not have sufficient evidence to reject the null hypothesis and so we accept the shipment.

It should be obvious that we can use $s^2$ as the test statistic in place of $\frac{(n-1)^2 s^2}{\sigma^2}$. If we were to use $s^2$ as the test statistic then, as before, we can reject the null hypothesis only when

$$\frac{(\mathrm{n}-1)\mathrm{s}^2}{\sigma^2} \geqslant 15.507, \text{ when } \sigma^2 = 0.25$$

i.e.$\frac{(9-1)s^2}{0.25} \geqslant 15.507$

i.e.$s^2 \geqslant 15.507 \times \frac{0.25}{8}$

or $s^2 \geqslant 0.485$

As our observed value of $s^2$ is only 0.32, we come to the same conclusion that the sample evidence is not strong enough for us to reject Ho.

**Two-Tailed Tests of Variance**

We have earlier used both one-tailed and two-tailed tests while discussing tests concerning population means and proportions. Similarly, depending on the situation, one may have to use a two-tailed test while testing for population variance.

The surface hardness of composite metal sheets is known to have a variance of 0.40. For a shipment just received, the sample variance from a random sample of nine sheets worked out to 0.22. Is it right to conclude that this shipment has a variance different from 0.40, if the significance level used is 0.05?

We start by stating our null and alternative hypotheses as below.

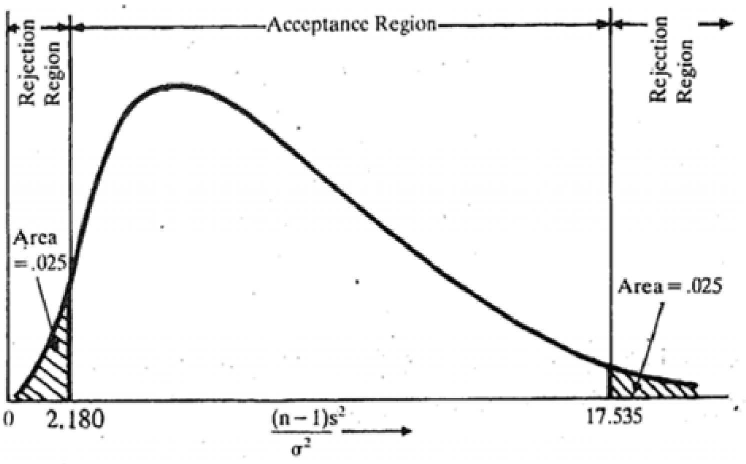
$$\mathrm{H}_0: \sigma^2 = 0.40$$

$$\mathrm{H}_1: \sigma^2 \neq 0.40$$

We shall again use $\frac{(n-1)s^2}{\sigma^2}$ as our test statistic which will have a chi-square distribution with (n -1) degrees of freedom, assessing the surface hardness of individual sheets followed a normal distribution.

Now, we shall reject the null hypothesis if the observed value of the test statistic is too small or too large. As the significance level of the test is 0.05, the probability of rejecting Ho when Ho is true is 0.05. Splitting this probability into two equal halves, we again have two critical regions each with an equal area as shown in Figure III below.

**Figure III: Acceptance and rejection regions for a two-tailed Test of Variance**



The test could, therefore, be summarised as follows:

Reject $H_0$, if $\frac{(n-1)s^2}{\sigma^2}$ is larger than 17.535 or if $\frac{(n-1)s^2}{\sigma^2}$ is smaller than 2.180, when $\sigma^2 = 0.40$ and n = 0. In other words,

Reject $H_0$, if $\frac{8s^2}{0.40}$ is larger than 17.535, or if $\frac{8s^2}{0.40}$ is smaller than 2.180/

The observed value of $s^2$ is 0.22 and so,

the observed value of $\frac{(n-1)s^2}{\sigma^2} = \frac{8\times0.22}{0.40} = 4.40$

As this value falls in the acceptance region of Figure III, the null hypothesis cannot be rejected and so we conclude that at a significance level of 0.05, there is not enough evidence to say that the variance of the shipment just received is different from 0.40.

**Activity A**

A psychologist is aware that the variability of attention-spans of five-year-olds can be minimised by $\sigma^2 = 49\ (minutes)^2$. While studying the attention-spans of 19 four-year- olds, it was found that $S^2 = 30$ (minutes)$^2$.

a) If you want to test whether the variability of attention-spans of the four-year-olds is different from that of the five-year-olds, what would be your null and alternative hypotheses?

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

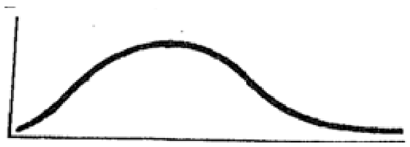
b) On the other hand, if you believe that the variability of attention-spans of the four-year-olds is not smaller than that of the five-year-olds, what would be your null and alternative hypotheses?

c) What test statistic would you choose for each of the above situations and what is the distribution of the test statistic that can be used to define the critical region?

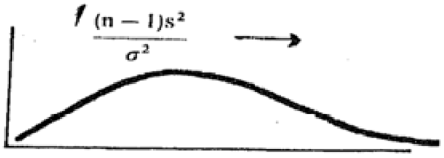
**Activity B**

For each of the folio wing situations, show the critical regions symbolically on the chi-square distributions shown alongside:

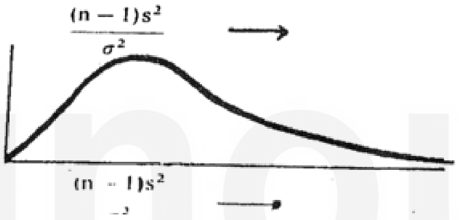
a) $\mathrm{H_0}: \sigma^2 \leqslant 0.5$
$H_1: \sigma^2 > 0.5$



b) $H_0: \sigma^2 = 0.5$
$H_1: \sigma^2 \neq 0.5$



c) $\mathrm{H_0}: \sigma_2^2 \geqslant 0.5$
$\mathrm{H_1}: \sigma_2^2 < 0.5$



## 12.3 TESTING OF EQUALITY OF TWO POPULATION VARIANCES

In many situations we might be interested in comparing the variances of the populations to see whether one is larger than the other or they are equal. For example, while testing the difference of means of two populations based on small independent samples in section 15.6 of the previous unit, we had assumed that both the populations had the same variance. We may want to test if it is reasonable to assume that the two population variances are equal.

While testing the equality of two population means, the test statistic used was the difference in two sample means. As we shall discover soon, while testing the equality of two population variances, the test statistic would be the ratio of the two sample variances.

### The F Distribution

If $\mathbf{X}_\mathrm{i}$ and $x_2$ are independent random variables having chi-square distributions with $v_1$ and $v_2$ degrees of freedom, then

$$y = \frac{x_1/\nu_1}{x_2/\nu_2}$$

has an F distribution with $\mathrm{V_i}$ and $v_2$ degrees of freedom.

The F distribution is also tabulated extensively and finds a lot of applications in applied statistics. An F distribution has two parameters-the first parameter refers to the degrees of freedom of the numerator chi-square random variable

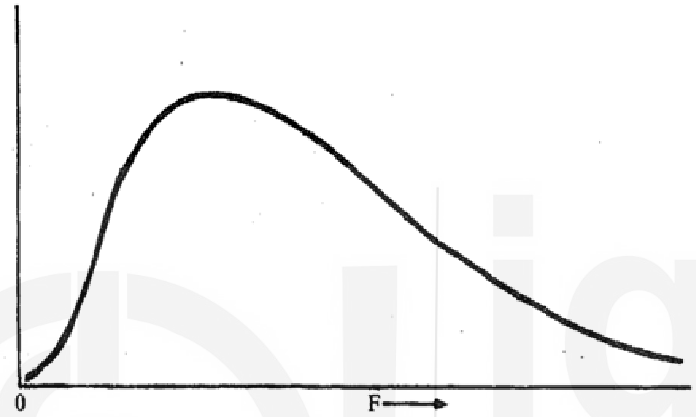
ignou
THE PEOPLE'S
UNIVERSITY

and the second parameter refers to the degrees of freedom of the denominator chi-square random variable.

The right tail of various F distributions with different numerator and denominator degrees of freedom is extensively tabulated. As we shall see later, the left tail of any F distribution can be easily calculated by some simple modifications.

Being a ratio of two chi-square variables (and their degrees of freedom), an F distribution exists for only positive values of the random variable. It is as symmetric and unimodal as shown in Figure IV below.

**Figure IV: An F distribution with $v_1$ and $v_2$ degrees of freedom (df)**



## A One-Tailed Test of Two Variances

A purchase manager wanted to test if the variance of prices of unbranded bolts was higher than the variance of prices of branded bolts. He needed strong evidence before he could conclude that the variance of prices of unbranded bolts was higher than the variance of prices of a branded bolts. He obtained price quotations from various stores and found that the sample variance of prices of unbranded bolts from 13 stores was 27.5. Similarly, the sample variance of prices of a certain brand of bolts from 9 stores was 11.2. What can the purchase manager conclude at a significance level of Let us use the subscript 1 for the population of prices of unbranded bolts and the subscript 2 for the population of prices of the given brand of bolts. We also assume that both these populations are normal. The purchase manager would conclude that the unbranded bolts have a higher price variance only when there was a strong evidence for it and not otherwise. So, the null and the alternative hypotheses would be:

$$H_0: \sigma_1^2 \leqslant \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

What should be the test statistic for this test? While testing the equality of two population means we had used the difference in sample means as the test statistic because the distribution of $(x_1 - x_2)$ and known. However, the distribution of $(s_1^2 - s_2^2)$ is not known and so this cannot be used as the test statistic. Let us see if we can know the distribution of $\frac{s_1^2}{s_2^2}$ when Ho is true.

Actually, we are interested in the distribution of the test statistic to define the critical region. The probability of type I error should not exceed the significance level, $\alpha$. This probability is the highest at the breakpoint between Ho and $H_1$, i.e. when $\alpha_1^2 = \alpha_2^2$ in this case.

Now, if both the populations are normal, then $\frac{(n_1-1)s_1^2}{\sigma_1^2}$ has a chi-square distribution with $(n_1 - 1)$ degrees of freedom, and $\frac{(n_2-1)s_2^2}{\sigma_2^2}$ has a chi-square distribution with $(n_2 - 1)$ degrees of freedom. These two samples can also be assumed to be independent and so

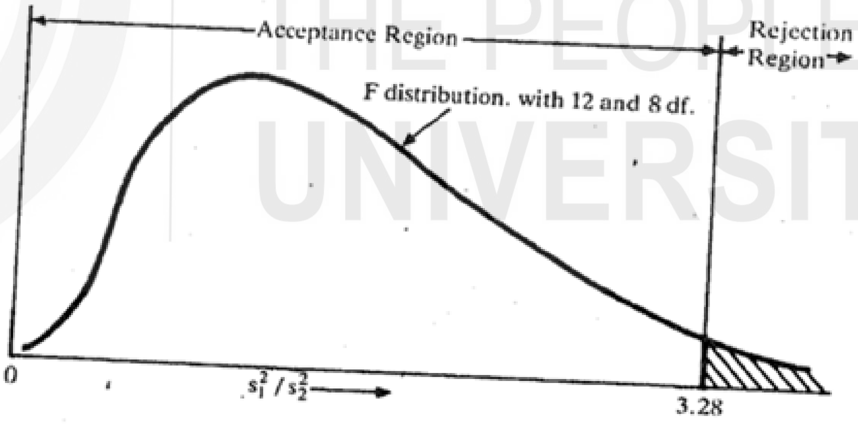
$$\frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)}$$

will have an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. But,

if $\sigma_1^2 = \sigma_2^2$ then, $\frac{s_1^2}{s_2^2}$ Will have an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

In this case $n_1 = 13, s_1^2 = 27.5; n_2 = 9, s_2^2 = 11.2$; and so by referring to the F tables for the distribution with 12 and 8 degrees of freedom, we find that the cut-off value of $\frac{s_1^2}{s_2^2}$ is 3.28, as shown in Figure V below.

**Figure V: Acceptance and Rejection Regions for a One-tailed Test of Equality of variance**



The observed value of $\frac{s_1^2}{s_2^2} = 27.5/11.2 = 2.455$

As this falls in the acceptance region of Figure V, we cannot reject Ho. Therefore, we conclude that we do not have sufficient evidence to justify that unbranded bolts have a higher price variance than that of a given brand.

### A Two-Tailed Test of Two Variances

A two-tailed test of equality of two variances is similar to the one-tailed test discussed in the previous section. The only difference is that the critical region would now be split into two parts under both the tails of the F distribution.

Let us take up the decision problem faced by the marketing manager in section 15.6 of the previous unit with some slightly different figures. Here the marketing manager wanted to know if display at point of purchase helped in increasing sales. He picked up 13 retail shops with no display and found that the weekly sale in these shops had a mean of Rs. 6,000 and a standard deviation of Rs. 1004. Similarly, he picked up a second sample of 11 retail shops with display at point of purchase and found that the weekly sale in these shops had a mean of Rs. 6500 and a standard deviation of Rs. 1,200. If he knew that the weekly sale in shops followed normal distributions, could he reasonably assume that the variances of weekly sale in shops with and without display were equal, if he used a significance level of 0.10?

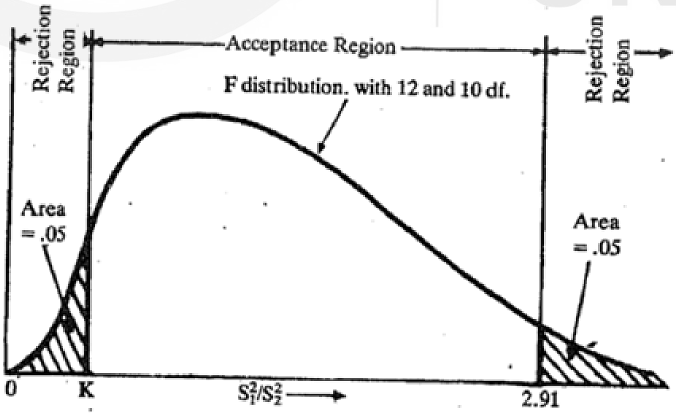
In section 15.1 we developed a test procedure based on the assumption that $\sigma_1 = \sigma_2$. Now we are interested in testing if that assumption is sustainable or not. We take the position that unless and until the evidence from the samples is strongly to the contrary we would believe that the two populations-viz. of shops without display and of shops with display-have equal variances. If we use the subscript 1 to refer to the former population and subscript 2 for the latter, then it follows that

$$\mathrm{H}_0: ;\sigma_1^2 = \sigma_2^2$$

$$\mathrm{H}_1: ;\sigma_1^2 \neq \sigma_2^2$$

We shall again use $\frac{s_1^2}{s_2^2}$ as the test statistic, which follows an F distribution with $s_2(\mathrm{n}_1 - 1)$ and $(\mathrm{n}_2 - 1)$ degrees of freedom, if the null hypothesis is true. This being a two¬tailed test, the critical region is split into two parts and as shown in Figure VI below, the upper cut-off point can be easily read off from the F tables as 2.91.

**Figure VI: Acceptance and Rejection Regions for a Two-tailed Test of Equality of Variances**



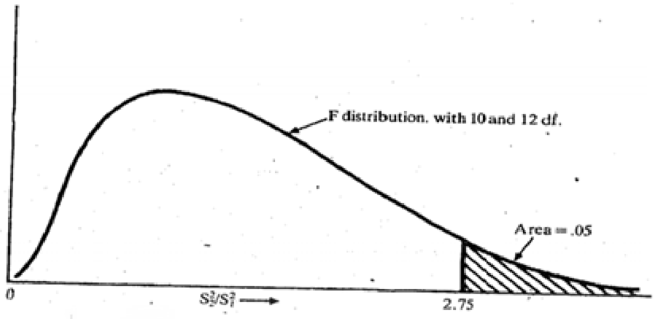
The lower cut-off point has been shown as K in Figure VI above and its value cannot be read off directly because the left tails of F distributions are not generally tabulated. However we know that K is such that

$$\Pr\left\{\frac{s_1^2}{s_2^2} \leqslant K\right\} = .05$$

i.e.Pr $\left\{\frac{s_2^2}{s_1^2} \geqslant 1/K\right\} = .05$

Now, $\frac{s_2^2}{s_1^2}$ will also have a F distribution with $(n_2 - 1)$ and $(n_1 - 1)$ degrees of freedom and so the value of 1/K can be easily looked up from the right tail of this 'distribution. As can be seen from Figure VII below, 1/K is equal to 2.75 and so K =1/2.75 = 0.363.

**Figure VII: The distribution of $S_2^2/S_1^2$**



Hence, the lower cut-off point for $\frac{s_1^2}{s_2^2}$ is 0.363: In other words, if the significance level is 0.10, the value of should lie between 0.363 and 2.91 for us to accept Ho. As the observed value of $\frac{s_1^2}{s_2^2} = \frac{1004^2}{1200^2} = 0.700$ which lies in the acceptance region, we accept the null hypothesis that variance from both populations are equal.

**Activity C**

From a sample of 16 observations, we find $S_1^2 = 3.52$ and from another sample of 13 observations, we find $S_2^2 = 4.69$. Under the assumption that $\sigma_1^2 = \sigma_2^2$, we find the following probabilities

$$\Pr\left\{-\frac{s_1^2}{s_2^2} \geqslant 2.62\right\} = .05$$

and Pr $\left\{\frac{s_2^2}{s_1^2} \geqslant 2.48\right\} = .05$

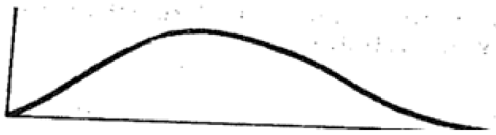
Find C such that

$$\Pr\left\{\frac{s_1^2}{s_2^2} \leqslant C\right\} = .05$$

**Activity D**

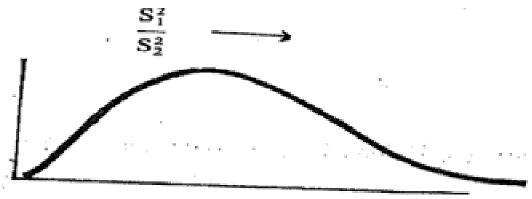
For each of the following situations, show the critical regions symbolically on the F distributions shown alongside:

a) $H_0: \sigma_1^2 \geqslant \sigma_2^2$
$H_1: \sigma_1^2 < \sigma_2^2$

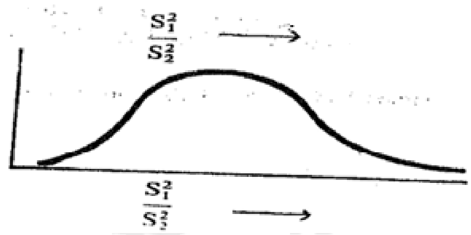


ignou
THE PEOPLE'S
UNIVERSITY

b) $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$



c) $H_0: \sigma_1^2 \leqslant \sigma_2^2$
$H_1: \sigma_1^2 > \sigma_2^2$



## 12.4 TESTING THE GOODNESS OF FIT

Many times we are interested in knowing. if it is reasonable to assume that the population distribution is Normal, Poisson, Uniform or any other known distribution. Again, the conclusion is to be based on the evidence produced by a sample. Such a procedure is developed to test how close is the fit between the observed data and the distribution assumed. These tests are also based on the chi-square statistic and we shall first provide a little background before such tests are taken up for detailed discussion.

**The Chi-Square Statistic**

Let us define a multinomial experiment which can be readily seen as an extension of the binomial experiment introduced in a previous unit. The experiment consists of making n trials. The trials are independent and the outcome of each trial falls into one of k categories. The probability that the outcome of any trial falls in a particular category, say category i, is $p_i$ and this probability remains the same from one trial to another. Let us denote the number of trials in which the outcome falls in category i by $n_1$. As the total number of trials is n and there are k categories in all, obviously

$$n_1 + n_2 + \cdots + n_k = n$$

Each one of the n, s is a random variable and their values depend on the outcome of n successive trials. Extending the concept from a binomial distribution, it is not difficult to see that the expected number of trials in which the outcome falls in category i, would be

E(ni) = n.pi, i = 1,2,... ,k

Now suppose that we hypothesis values for $p_1, p_2, \ldots, p_k$. If the hypothesis is true then the observed values of n would not be greatly different from the expected number in category is the random variable $\chi^2$ defined as below, will approximately possess a chi-square distribution.

$$\chi^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{k} \frac{[n_i - np_i]^2}{E(n_i)}$$

It is easy to see that when there are only two categories (i.e. k = 2), we will approximately have a chi-square distribution. In such a case $p_1 + p_2 = 1$ and so

$$\chi^2 = \frac{(n_1-np_1)^2}{n_{P_1}} + \frac{(n_2-np_2)^2}{np_2}$$

$$= \frac{(n_1 - np_1)^2 \cdot p_2 + (n_2 - np_2)^2 \cdot p_1}{np_1 p_2}$$

$$= \frac{(n_1 - np_1)^2 \cdot p_2 + [(n - n_1) - n(1 - p_1)]^2 \cdot p_1}{np_1(1 - p_1)}$$

$$= \frac{(n_1 - np_1)^2 \cdot P_2 + (-n_1 + np_1)^2 \cdot p_1}{np_1(1 - p_1)}$$

$$= \frac{(n_1 - np_1)^2}{np_1(1 - p_1)}$$

But from our earlier discussion of the normal approximation to the binomial distribution, we know that when n is large, $\frac{n_1-np_1}{\sqrt{np_1(1-p_1)}}$ has a standard normal distribution and so $x^2$ above will have a chi-square distribution with one degree of freedom.

In general, when the number of categories is K, $X^2$ has a chi-square distribution with (k - 1) degrees of freedom. One degree of freedom is lost because of one linear constraint on the $n_i$ 's, viz.

$$n_1 + n_2 + \cdots + n_k = n$$

The $\chi^2$ statistic would approximately have a chi-square distribution when n is sufficiently large so that for each i, $np_i$ is at least 5-i.e. the expected frequency in each category is at least equal to 5.

Using a different set of symbols, if we write $O_i$ or the observed frequency in category i and Ei for the expected frequency in the same category, then the chi-square statistic and also be computed as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

**An Example: Testing for Uniform Distribution**

Suppose we want to test if a worker is equally prone to producing defective components throughout an eight hour shift or not. We break the shift into four two- hour slots and count the number of defective components produced in each of these slots. At the end of one week we find that the worker has produced 50 defective components with the following break-up:

| **Time Slot (hours)** | **Observed Frequency** |
|---|---|
| 8.00-10.00 | 8 |
| 10.00-12.00 | 11 |
| 12.30-14.30 | 16 |
| 14.30-46.30 | 15 |
|  | 50 |

ignou THE PEOPLE'S UNIVERSITY

From this data using a significance level of .05, is it reasonable to assume that the probability to produce a defective component is equal in each of the four two-hour slots?

We shall take the position that unless and until the sample evidence is overwhelmingly against it, we shall accept that the probability to produce a defective component in any two-hour slot is the same. If we represent the probability that a defective component came from the $i^{th}$ slot by $p_i$, then the null and the alternative hypotheses are:

$$H_0: p_1 = p_2 = p_3 = p_4 = 0.25$$

$H_1$ All of $p_1, p_2, p_3$ and $p_4$ are not equal.

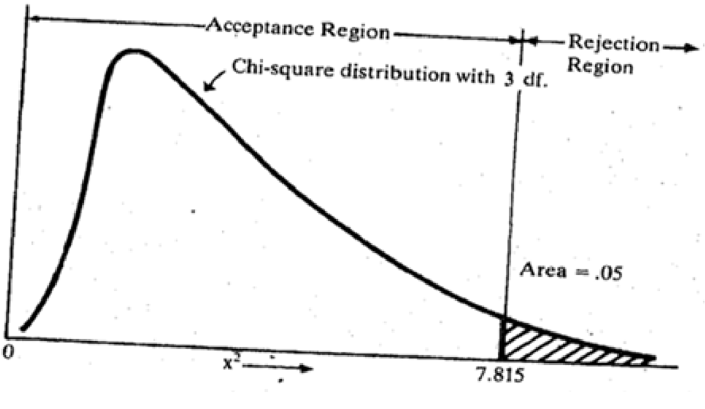
We shall use the chi-square statistic $x^2$ as our test statistic and the expected frequencies would be computed based on the assumption that the null hypothesis is true. This and some more computations have been made in Table 1 below.

**Table 1: Computation of the Chi-Square Statistic**

| S1.No. (i) | Time Slot (hours) | Obs. Freq. (Oi) | Exp. Freq. ($E_i$) | $O_i - E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|---|---|
| 1 | 8.00-10.00 | 8 | 12.50' | - 4.50 | 20.25 | 1.62 |
| 2 | 10.00-12.00 | 11 | 12.50 | - 1.50 | 2.25 | 0.18 |
| 3 | 12.30-14:30 | 16 | 12.50 | 3.50 | 12.25 | 0.98 |
| 4 | 14.30-16.30 | 15 | IH12.50 | 2.50 | 6.25 | 0.50 |
|  | Total | 50 | 50.00 |  |  | 3.28 |

In the above table, the expected frequencies E have been calculated as $np_i$ where n, the total frequency is 50 and each $p_i$ is 0.25 under the null hypothesis. Now, if the null hypothesis is true, $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ will have a chi-square distribution with (k -1), i.e. (4 - 1) = 3 degrees of freedom and so if we want a significance level of .05,then as shown in Figure VIII below, the cut-off value of the chi-square statistic should be 7.815.

**Figure VIII: Acceptance and Rejection Regions for a .05 significance level Test**



ignou THE PEOPLE'S UNIVERSITY

Therefore, we can reject the null hypothesis only when the observed value of the chi¬square statistic is at least 7.815. As the observed value of the chi-square statistic is only 3.28, we cannot reject the null hypothesis.

Using the concepts developed so far, it is not difficult to see how a test procedure can be developed and used to test if the data observed came from any known distribution. The degrees of freedom for the chi-square statistic would be equal to the number of categories (k) minus 1 minus the number of independent parameters of the distribution estimated from the data itself.

If we want to test whether it is reasonable to assume that an observed sample came from a normal population, we may have to estimate the mean and the variance of the normal distribution first, We would categorise the observed data into an appropriate number of classes and for each class we would then calculate the probability that the random variable belonged to this class, if the population distribution were normal. Then, we would repeat the computations as shown in this section-viz. calculating the expected frequency in each class. Finally, the value of chi-square statistic would have (k - 3) degree of freedom since two parameters (the mean and the variance) of the population were estimated from the sample.

**Activity E**

From the following data, test if it is reasonable to assume that the population has a distribution with $p_i = 0.2, p_2 = 0.3$ and $p_3 = 0.5$. Use α = .05.

| Category (i) | $O_i$ | $P_i$ | $E_i$ | $(O_i - E_i)$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|---|---|
| 1 | 17 | 0.2 | | $(O_i - E_i)$ | $(O_i - E_i)^2$ | |
| 2 | 35 | 0.3 | | | | |
| 3 | 48 | 0.5 | | | | |
| Total | 100 | 1.0 | | | | |

## 12.5 TESTING INDEPENDENCE OF CATEGORISED DATA

A problem frequently encountered in the analysis of categorised data concerns the independence of two methods of classification of the observed data. For example, in a survey, the responding consumers could be classified according to their sex and their preference of our product over the next competing brand-(again measured by classifying them into three categories of preference). Such data is first prepared in the form of a contingency (or dependency) table which helps in the investigation of dependency between the classification criteria.

We want to study if the preference of a consumer for our brand and shampoo depends on his or her income level using a significance level of .05. We survey a total of 350 consumers and each is classified into (1) one of three income levels defined by us and (2) one of four categories of preference for our brand of shampoo over the next competing brand-viz., 'strongly prefer', 'moderately prefer', indifferent' and 'do not prefer'. These observations are presented in the form of a contingency table in Table 2 below.

The table shows, for example, that out of 350 consumers observed 98 belonged to the high income category, 108 to the medium income category and 144 to the low income group. Similarly, there were 95 consumers who strongly preferred our brand, 119 who moderately preferred our brand and so on. Further, the contingency table tells us that 15 consumers were observed to belong to both the high income level and the "strongly prefer" category of preference, and so on for the rest of the cells.

**Table 2: Observed (Expected) Frequencies in a Contingency Table**

| | **Category of Preference** | | | | |
|---|---|---|---|---|---|
| **Income Level** | **Strongly Prefer** | **Moderately Prefer** | **Indifferent** | **Do not Prefer** | **Total** |
| High | 15 (26.60) | 35 (33.32) | 21 (16.52) | 27 (21.56) | 98 |
| Medium | 48 (29.31) | 18 (36.72) | 20 (18.21) | 22 (23.76) | 108 |
| Low | 32 (39.09) | 66 (48.96) | 18 (24.27) | 28 (31.68) | 144 |
| Total | 95 | 119 | 59 | 77 | 350 |

Let $p_i$ = marginal probability for the $i^{th}$ row, i = 1, 2,…, r where r is the total number of rows. In this case $p_i$ would mean the probability that randomly selected consumer would belong to the $i^{th}$ income level.

P = marginal probability for the jth column, j = 1, 2, ... c, where c is the total number of columns. In this case $p_i$ would mean that the probability that a randomly selected consumer would belong to the $j^{th}$ preference category.

and $p_{ij}$ = Joint probability for the $i^{th}$ row and the $j^{th}$ column. In this case $p_{ij}$ would refer to the probability that a randomly selected consumer belongs to the $i^{th}$ income level and the $j^{th}$ preference category.

Now we can state our null and the alternative hypotheses as follows:

Ho: the criteria for column classification is independent of the criteria for row classification.

In this case, this would mean that the preference for our brand is not independent of the income level of the consumers.

$H_1$: the criteria for column classification is not independent of the criteria for row classification.

If the row and the column classifications are independent of each other, then it would follow that $p_{ij} = p_i \times p_j$

This can be used to state our null and the alternative hypotheses:

$$H_0: p_{ij} = p_i \times p_{.j} \text{ for } i = 1,2,\dots,r \text{ and } j = 1,2,\dots,c$$

$$H_1: p_{ij} \neq p_i \times p_{.j} \text{ for } i = 1,2,\dots,r \text{ and } j = 1,2,\dots,c$$

Now we know how the test has to be developed. If $p_i$ and $p_j$ are known, we can find the probability and consequently the expected frequency in each of the (r x c) cells of our contingency table and from the observed and the expected frequencies, compute the chi-square statistic to conduct the test. However, since the $p_i$'s and $p_j$'s are known, we have to estimate these from the data itself.

If $n_i$ = row total for the $i^{th}$ row

$n_j$ = column total for the $\mathrm{j}^{th}$ column

and n = the total of all observed frequencies.

then our estimate of $\mathrm{p}_{i,} = \mathrm{n_i}/\mathrm{n}$

and our estimate of $\mathrm{p_j} = \mathrm{n_j}/\mathrm{n}$

and so the expected frequency in the $i^{th}$ row and column $j^{th}$

$$E_{\mathrm{ij}} = \mathrm{np_{ij}} = \mathrm{n(p_{i.})(p_{.j})} = \mathrm{n} \times \mathrm{(n_i/n)x(n_j/n)} = \mathrm{(n_i x n_j)/n}$$

and if the observed frequency in the $i^{th}$ and column is referred to as $\mathrm{O_{ij}}$ then the chi¬square statistic can be computed as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

This statistic will have a chi-square distribution with the degrees of freedom given by the total number of categories or cells (i.e. r x c) minus 1 minus the number of independent parameters estimated from the data. We have estimated r marginal row probabilities out of which (r - 1) have been independent, since

$$p_1 + p_2 + \cdots + p_r = 1$$

Similarly, we have estimated c marginal column probabilities out of which (c - 1) have been independent, since

$$p_1 + p_2 + \cdots \dots \dots + +p_c = 1$$

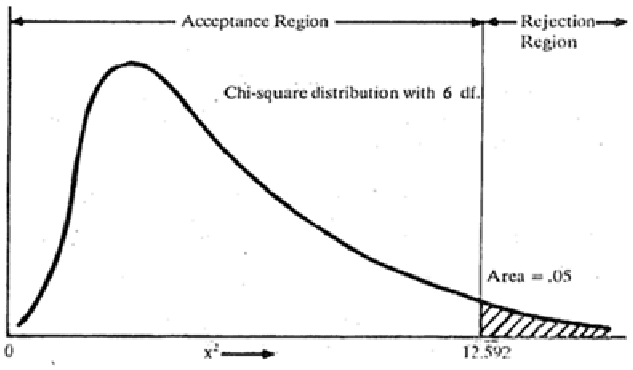
and so, the degrees of freedom for the chi-square statistic

$$= rc - 1 - (r - 1) - (c - 1)$$

$$= (r - 1)(c - 1)$$

Coming back to the problem at hand, the chi-square statistic computed as will have (3-1) (4-1) i.e. 6 degrees of freedom and so by referring to the Figure IX below, we can say that we would reject the null hypothesis at a significance level of 0.05, if the computed value of $x^2$ above is greater than or equal to 12.592.

Figure IX: Rejection region for a test using the Chi-square statistics



Now, the only task is to compute the value of the chi-square statistic. For this, we first find the expected frequency in each cell using the relationship.

$$E_{ij} = \frac{n_i \times n_j}{n}$$

For example, when i = 1 and j = 1, we find

$$\mathrm{F}_{11} = \frac{98 \times 95}{350} = 26.60$$

These values have also been recorded in Table 2 in parentheses and so the chi-square statistic is computed as

$$x^2 = \frac{(15-26.60)^2}{26.60} + \frac{(35-33.32)^2}{33.32} + \frac{(21-16.52)^2}{16.52} + \frac{(27-21.56)^2}{21.56}$$

$$+ \frac{(48-29.31)^2}{29.31} + \frac{(18-36.72)^2}{36.72} + \frac{(20-18.21)^2}{18.21} + \frac{(22-23.76)^2}{23.76}$$

$$\frac{(32-39.09)^2}{39.09} + \frac{(66-48.96)^2}{48.96} + \frac{(18-24.27)^2}{24.27} + \frac{(28-31.68)^2}{31.68}$$

$$= 5.059 + 0.085 + 1.215 + 1.373 + 11.918 + 9.544 + 0.176 + 0.130 + 1.286$$

$$+5.930 + 1.620 + 0.427$$

= 38.763

As the computed value of the chi-square statistic is much above the cut-off value of 12.592, we reject the null hypotheses at a significance level of 0.05 and conclude that the income level and preference for our brand are not independent.

Whenever we are using the chi-square statistic we must make sure that there are enough observations so that the expected frequency in any cell is not less than 5; if not, we may have to combine rows or columns to raise the expected frequency in each cell to at least 5.

## 12.6 SUMMARY

In this unit we have looked at some situations where we can develop tests based on the chi-square distribution. We started by testing the variance of a normal population where the test statistic used was $\frac{(n-1)s^2}{\sigma^2}$ since the distribution of the sample variance $s^2$ was not known directly. We found that such tests could be one-tailed depending on our null and the alternative hypotheses.

We then developed a procedure for testing the equality of variances of two normal populations. The test statistic used in this case was the ratio of the two sample variances are this was found to have a F distribution under the null hypothesis. This procedure enabled us to test the assumption made while we developed a test procedure for testing the equality of two population means based on a small independent samples in the previous unit.

We then described a multinomial experiment and found that if we have data that classify observations into k different categories and if the conditions for the multinomial experiment are satisfied then a test statistic called the chi-square statistic defined as $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ will have a chi-square distribution with specified degrees of freedom. Here, $O_i$ refers to the observed frequency of the $i^{th}$ category and $E_i$ to the expected frequency of the $i^{th}$ category and the degree of freedom is equal to the number of categories minus 1 minus the number of independent parameters estimated from the data to calculate the E's. This concept was used to develop tests concerning the goodness of fit of the observed data to any hypothesised distribution and also to test if two criteria for classification are independent or not.

## 12.7 SELF-ASSESSMENT EXERCISES

1) A production manager is certain that the output rate of experienced employees is better than that of the newly appointed employees. However, he is not sure if the variability in output rates for these two groups is same or not. From previous studies it is known that the mean output rate per hour of new employees at a particular work centre is 20 units with a standard deviation of 4 units. For a group of 15 employees with three year's experience, it was found that the sample mean of output rate per hour was 30 units with a sample standard deviation of 6 units. Is it reasonable to assume that the variability of output rates at these two experience levels is not different? Test at a significance level of .01.

2) For self-assessment exercise No. of the previous unit test if it is reasonable to assume $\sigma = \sigma_2$ at $\alpha$ = .05.

3) The safety manager of a large chemical plant went through the file of minor accidents in his plant and picked up a random sample of some accident and classified them according to the time at which the accident took place. Using the chi-square test at a significance level of 0.01. What should we conclude? If you were the safety manager, what would you do after completing the test?

| **Time (hrs.)** | **No. of Accidents** |
|---|---|
| 3.00-9.00 | 6 |

| | |
|---|---|
| 9.00-10.00 | 7 |
| 10.00-11.00 | 21 |
| 11.00-12.00 | 9 |
| 13.00-14.00 | 7 |
| 14.00-15.00 | 8 |
| 15.00-16.00 | 18 |
| 16.00-17.00 | 9 |

4) A survey of industrial sales persons included questions on the age of the respondent and the degree of job pressure the sales person felt in connection with the job. The data is presented in the table below. Using a significance level of .01, examine if there is any relationship between the age and the degree of job pressure.

| | **Degree of pressure job** | | |
|---|---|---|---|
| Age (years) | Law | Medium | High |
| Less than 25 | 32 | 25 | 17 |
| 25-34 | 22 | 19 | 20 |
| 35-54 | 17 | 20 | 25 |
| 55 and above | 15 | 24 | 26 |

For each of the statements below, choose the most appropriate response from among the ones listed.

5) The major reason that chi-square tests for independence and for goodness of fit are one-tailed is that:

a) small values of the test statistic provide support for Ho

b) large values of the test statistic provide support for Ho

c) tables are usually available for right-tailed rejection regions

d) none of the above.

6) When testing to draw inferences about one or two population variances, using the chi-square and the F distributions, respectively, the major assumption needed is

a) large sample sizes

b) equality of variances

c) normal distributions of population

d) all of the above.

7) In chi-square tests of goodness of fit and independence of categorical data, it is sometimes necessary to reduce the numbers of classifications used to

a) provide the table.with larger observed frequencies

b) make the distribution appear more normal

c) satisfy the condition that variances must be equal

ignou
THE PEOPLE'S
UNIVERSITY

d) none of the above.

8) In carrying out a chi-square test of independence of categorical data, we use all of the following except

   a) an estimate of the population variance

   b) contingency tables

   c) observed and expected frequencies

   d) number of rows and columns.

9) The chi-square distribution is used to test a number of different hypotheses. Which of the following is an application of the chi-square test?

   a) goodness-of-fit of a distribution

   b) equality of populations

   c) Independence of two variables or attributes

   d) all of the above.

## 12.8 FURTHER READINGS

Bruce Bomerman, Business Statistics for Practice, McGraw Hill.

Gravetter F.J. and L.B. Wallrnce. *Statistics for the Behavioural Sciences*, West Publishing Co.: St. Paul.

Minnesota Levin R.I., *Statistics for Management*: 1 5rentice-Hall of India: New Delhi.

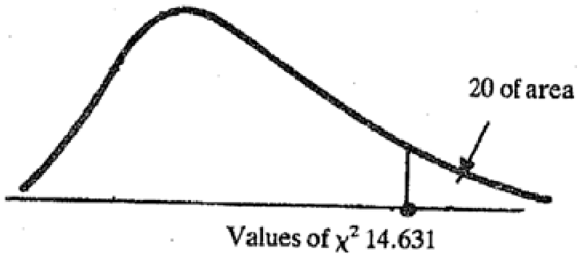
Mason R.D., *Statistical Techniques in Business and Economics*, Richard D. Irwin, Inc: Homewood, Illinois.

Mendenhall W., Schaffer R.L. and D.D. Wackerly. *Mathematical Statistics with Applications,* Duxbury Press: Boston Monachasetts.

Plane D .R. and E.B. Oppern,ann. *Business and Economic Statistics*, Business Publications, Inc: Plano, Texas.

**APPENDIX TABLE 5**

Area in the Right Tail of a Chi-square ($x^2$) Distribution.[1]



---

[1] Taken from Table IV of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London (previously published by Oliver & Boyd, Edinburgh and by premission of the authors and publishers.

| Degrees of freedom | 99 | 0.975 | 0.95 | 0.90 | 0.800 |
|---|---|---|---|---|---|
| 1 | 0.00016 | 0.00098 | 0.00398 | 0.0158 | 0.642 |
| 2 | 0.0201 | 0.0506 | 0.103 | 0.211 | 0.446 |
| 3 | 0.115 | 0.216 | 0.352 | 0.584 | 1.005 |
| 4 | 0.297 | 0.484 | 0.711 | 1.064 | 1.649 |
| 5 | 0.554 | 0.831 | 1.145 | 1.610 | 2.343 |
| 6 | 0.872 | 1.237 | 1.635 | 2.204 | 3.070 |
| 7 | 1.239 | 1.690 | 2.167 | 2.833 | 3.822 |
| 8 | 1.646 | 2.180 | 2.733 | 3.490 | 4.594 |
| 9 | 2.088 | 2.700 | 3.325 | 4.168 | 5.380 |
| 10 | 2.558 | 3.247 | 3.940 | 4.865 | 6.179 |
| 12 | 3.571 | 4.404 | 5.228 | 6.304 | 7.807 |
| 13 | 4.107 | 5.009 | 5.892 | 7.042 | 8.634 |
| 14 | 4.660 | 5.629 | 6.571 | 7.790 | 9.467 |
| 15 | 5.229 | 6.262 | 7.261 | 8.547 | 10.307 |
| 16 | 5.812 | 6.908 | 7.962 | 9.312 | 11.152 |
| 17 | 6.408 | 7.564 | 8.672 | 10.085 | 12.002 |
| 18 | 7.015 | 8.231 | 9.390 | 10.865 | 12.587 |
| 19 | 7.633 | 8.907 | 10.117 | 11.851 | 13.716 |
| 20 | 8.260 | 9.591 | 10.851 | 12.443 | 14.578 |
| 21 | 8.897 | 10.283 | 11.591 | 13.240 | 15.445 |
| 22 | 9.542 | 10.982 | 12.338 | 14.041 | 16.314 |
| 23 | 10.196 | 11.889 | 13.091 | 14.848 | 17.187 |
| 24 | 10.856 | 12.401 | 13.848 | 15.658 | 18.062 |
| 25 | 11.524 | 13.120 | 14.611 | 16.473 | 18.940 |
| 26 | 12.198 | 13.844 | 15.879 | 17.292 | 19.820 |
| 27 | 12.879 | 14.573 | 16.151 | 18.114 | 20.703 |
| 28 | 13.565 | 15.308 | 16.928 | 18.939 | 21.588 |
| 29 | 14.256 | 16.047 | 17.708 | 19.768 | 22.475 |
| 30 | 14.953 | 16.791 | 18.493 | 20.599 | 23.364 |

* Taken from Table IV of Fisher and Yates, Statistical Tables for Biological, Agricultural and Medical Rsearch, Published by Longman Group Ltd., London (previously published by Oliver & Boyd, Edinburgh and by permission of the authors and publishers.

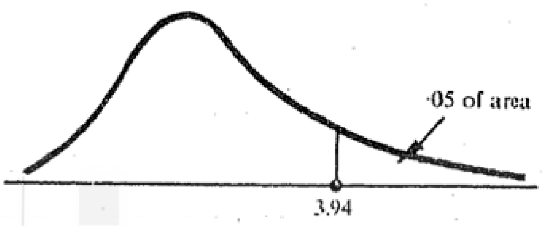
| 0.20 | .10 | .05 | 0.025 | .01 | Degrees of freedom |
|---|---|---|---|---|---|
| 1.642 | 2.706 | 3.841 | 5.024 | 6.635 | 1 |
| 3.219 | 4.605 | 5.991 | 7.378 | 9.210 | 2 |
| 4.642 | 6.251 | 7.815 | 9.48 | 11.345 | 3 |
| 5.989 | 7.779 | 9.488 | 11.143 | 13.277 | 4 |
| 7.289 | 9.236 | 11.070 | 12.833 | 15.086 | 5 |
| 8.558 | 10.645 | 12.592 | 14.449 | 16.812 | 6 |
| 9.803 | 12.017 | 14.067 | 16.013 | 18.475 | 7 |
| 11.030 | 13.362 | 15.507 | 17.535 | 20.090 | 8 |
| 12.242 | 14.684 | 16.919 | 19.023 | 21.666 | 9 |
| 13.442 | 15.987 | 18.307 | 20.483 | 23.209 | 10 |
| 14.631 | 17.275 | 19.675 | 21.920 | 24.725 | 11 |
| 15.812 | 18.549 | 21.026 | 23.337 | 26.217 | 12 |
| 16.985 | 19812 | 22.362 | 24.736 | 27.688 | 13 |
| 18.151 | 21.064 | 23.685 | 26.119 | 29.141 | 14 |
| 19.311 | 22.307 | 24.996 | 27.488 | 30.578 | 15 |
| 20.465 | 23.542 | 26.296 | 28.845 | 32.000 | 16 |
| 21.615 | 24.769 | 27.587 | 30.191 | 33.409 | 17 |
| 22.760 | 25.989 | 28.869 | 31.526 | 34.805 | 18 |
| 23.900 | 27.204 | 30.144 | 32.852 | 36.191 | 19 |
| 25.038 | 28.412 | 31.410 | 34.170 | 37.566 | 20 |

ignou THE PEOPLE'S UNIVERSITY

| | | | | | |
|---|---|---|---|---|---|
| 26.171 | 29.615 | 32.671 | 35.479 | 38.932 | 21 |
| 27.301 | 30.813 | 33.924 | 36.781 | 40.289 | 22 |
| 28.429 | 32.007 | 35.172 | 38.076 | 41.638 | 23 |
| 29.553 | 33.196 | 36.415 | 39.364 | 42.980 | 24 |
| 30.675 | 34.382 | 37.652 | 40.647 | 44.314 | 25 |
| 31.795 | 35.563 | 38.885 | 41.923 | 45.642 | 26 |
| 32.912 | 36.741 | 40.113 | 43.194 | 46.963 | 27 |
| 34.027 | 37.916 | 41.337 | 44.461 | 48.278 | 28 |
| 35.139 | 39.087 | 42.557 | 45.722 | 49.588 | 29 |
| 36.250 | 40.256 | 43.773 | 46.979 | 50.892 | 30 |

## APPENDIX TABLE 6

Values of F for F Distributions with .05 of the Area in the Right Tail[2]



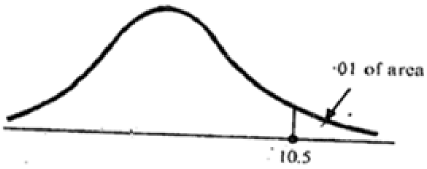
Degrees of freedom for numerator

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∝ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 246 | 2'48 | 249 | 250 | 251 | 252 | 253 | 254 |
| 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.5 | 19.5 | 19.5 | 19.5 | 19.5 | 19.5 |
| 3 | 10.1 | 9.55 | 9.28' | 9.12 | 9.01 | 8.94 | 8.89 | H.K5 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.65 | 8.53 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| 5 | 0.61 | 5.79 | 5.4! | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.37 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3,73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3,23 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01. | 2.97 | 2.93 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3,18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.91) | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2,57 | 2.53 | 2.49 | 2.45 | 2.40 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2,67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2,70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59.) | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 |
| 16 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.3.4 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2,48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2,42 | 2.37 | 2.32 | 2.25 | 2.18 | 2:10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2,53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2,36 | 2.30 | 2.25 | 2.18 | 2,11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 |
| 25 | 4.24 | 3.39 | 2,99 | 2.76 | 2.60 | 2.49 | 2.40 | 2,34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.2! | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2,12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.3'J |
| 20 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.83 | 1.75' | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 |
| cc | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1 22 | 1.00 |

**Value for F for Distribution with .01 of the Area in the Right Tai**

---

[2] Source: M. Mervin'ton and C.M. Thompson, *Riontetrika*, vol. 33 (1943).



**Degrees of freedom for numerator**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∝ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4,052 | 5,000 | 5,403 | 5,625 | 5,764 | 5,859 | 5,928 | 5,982 | 6,023 | 6,056 | 6,106 | 6,157 | 6,209 | 6,235 | 6,261 | 6,287 | 6,313 | 6,339 | 6,366 |
| 2 | 98.5 | 99.0 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 | 99.5 | 99.5 | 99.5 |
| 3 | 34.1 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.3 | 27.2 | 27.1 | 26.9 | 26.7 | 26.6 | 26.5 | 26.4 | 26.3 | 26.2 | 26.1 |
| 4 | 21.2 | 18.0 | 16.7 | 16.0 | 15.5 | 15.2 | 15.0 | 14.8 | 14.7 | 14.5 | 14.4 | 14.2 | 14.0 | 13.9 | 13.8 | 13.7 | 13.7 | 13.6 | 13.5 |
| 5 | 16.3 | 13.3 | 12.1 | 11.4 | 11.0 | 10.7 | 10.5 | 10.3 | 10.2 | 10.1 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.02 |
| 6 | 13.7 | 10.9 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.88 |
| 7 | 12.2 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.65 |
| 8 | 11.3 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.86 |
| 9 | 10.6 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.31 |
| 10 | 10.0 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.91 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.60 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.36 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 3.17 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.70 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.00 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.87 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.75 |
| 17 | 8.40 | 6.11 | 5.19 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.65 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.57 |
| 19 | 8.19 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.49 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.42 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 2.36 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.31 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.26 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 2.21 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.86 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.53 | 2.45 | 2.36 | 2.27 | 2.17 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.01 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.80 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.60 |
| 20 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 | 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.38 |
| ∞ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 | 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | 1.00 |

**Degrees of freedom for denominator.**

THE PEOPLE'S UNIVERSITY

# BLOCK 4

# FORECASTING METHODS

ignou
THE PEOPLE'S
UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

# UNIT 13 BUSINESS FORECASTING

## Objectives

After completion of this unit, you should be able to :

- realise that forecasting is a scientific discipline unlike ad hoc predictions
- appreciate that forecasting is essential for a variety of planning decisions
- become aware of forecasting methods for long, medium and short term decisions
- use Moving Averages and Exponential smoothing for demand forecasting
- understand the concept of forecast control
- use the moving range chart to monitor a forecasting system.

## Structure

13.1 Introduction

13.2 Forecasting for Long Term Decisions

13.3 Forecasting for Medium and Short Term Decisions

13.4 Forecast Control

13.5 Summary

13.6 Self-assessment Exercises

13.7 Key Words

13.8 Further Readings

ignou THE PEOPLE'S UNIVERSITY

## 13.1 INTRODUCTION

Data on demands of the market may be needed for a number of purposes to assist an organisation in its long term, medium and short term decisions. Forecasting is essential for a number of planning decisions and often provides a valuable input on which future operations of the business enterprise depend. Some of the areas where forecasts of future product demand would be useful are indicated below :

1) Specification of production targets as functions of time.
2) Planning equipment and manpower usage, as well as additional procurement.
3) Budget allocation depending on the level of production and sales.
4) Determination of the best inventory policy.
5) Decisions on expansion and major changes in production processes and methods.
6) Future trends of product development, diversification, scrapping etc.

7) Design of suitable pricing policy.

8) Planning the methods of distribution and sales promotion.

It is thus clear that the forecast of demand of a product serves as a vital input for a number of important decisions and it is, therefore, necessary, to adopt a systematic and rational methodology for generating reliable forecasts.

### The Uncertain Future

The future is inherently uncertain and since time immemorial man has made attempts to unravel the mystery of the future. In the past it was the crystal gazer or a person allegedly in possession of some supernatural powers who would make predications about the things-to be-major events or the rise and fall of kings. In today's world, predictions are being made daily in the realm of business, industry and politics. Since the operation of any capital enterprise has a large lead time (1-5 years is typical), it is clear that a factory conceived today is for some future demand and the whole operation is dependent on the actual demand coming up to the level projected much earlier. During this period many circumstances, which might not even have been imagined, could come up. For instance, there could be development of other industries, or a major technological breakthrough that may render the originally conceived product obsolete; or a social upheaval and change-of government may redefine priorities of growth and development; or an unusual weather condition like drought or floods may alter completely the buying potential of the originally conceived market. This is only a partial list to suggest how uncertainties from a variety of sources can enter to make the task of prediction of the future extremely difficult.

It is proper at this stage to emphasise the distinction between prediction and forecasting. Forecasting generally refers to the scientific methodology that often uses past data along with some well-defined assumptions or 'model' to come up with a 'forecast' of future demand. In that sense, forecasting is objective. A prediction is a subjective estimate made by an individual by using his intuitive 'hunch' which may in fact come out true. But the fact that it is subjective (A's prediction may be different from B's and C's) and non-realisable as a Well-documented computer programme (which could be used by anyone) deprives it of much value. This is not to discount. the role of intuition or subjectivity in practical decision-making. In fact, for complex long term decisions, intuitive methods such as the Delphi technique are most popular. The opinion of a well informed, educated person is likely to be reliable, reflecting the well-considered contribution of a host of complex factors in a relationship that may be difficult to explicitly quantify. Often forecasts are modified based on subjective judgment and experience to obtain predictions used in planning and decision making.

The future is inherently uncertain and any forecast at best is an educated guess with no guarantee of coming true. In certain purely deterministic systems (as for example in classical physics the laws governing the motion of celestial bodies are fairly well developed) an unequivocal relationship. between cause and effect has been clearly established and it is possible to predict. very accurately the course of events in the future, once the future patterns of causes are inferred from past behaviour. Economic systems, however, are more complex because (i) there is a large number of governing factors in a complex structural framework which may not be possible to identify and (ii) the individual factors themselves have a high degree of variability and uncertainty. The demand for a particular product (say umbrellas) would depend on competitor's prices, advertising campaigns, weather conditions, population and a number of factors which might even be difficult to identify. In spite of these complexities, a forecast has to be made so that the manufacturers of umbrellas (a product which exhibits a seasonal demand) can plan for the next season.

### Forecasting for Planning Decisions

The primary purpose of forecasting. is to provide valuable information for planning the design and operation of the enterprise. Planning decisions may be classified as long term, medium term and short term.

Long term decisions include decisions like plant expansion or. new product introduction which may require new technologies or a. complete transformation in social or moral fabric of society. Such decisions are generally, characterised by lack of quantitative information and absence of historical data on which to base, the forecast of future events. Intuition and the collected opinion of. experts in the field generally play a significant role in developing forecasts for such decisions. Some methods used in forecasting for long term decisions are discussed in Section 13.2.

Medium term decisions involve such decisions as planning the production levels in a manufacturing plant over the next year, determination of manpower requirements or inventory policy for the firm. Short term decisions include daily production planning and scheduling decisions. For both medium and short term forecasting, many methods and techniques exist. These methods can broadly be classified as follows

a) Subjective of intuitive methods.

b) Methods based on averaging of past data, including simple, weighted and moving averages.

c) Regression models on historical data.

d) Causal or Econometric models.

e) Time series analysis or stochastic models.

These methods are briefly reviewed in Section 13.3. A more detailed discussion of correlation, regression and time series models is taken up in the next three units.

The choice of an appropriate forecasting method is discussed in Section 13.4. The aspect of forecast control which tells whether a particular method in use is acceptable is discussed in Section 13.5. And finally a summary is given in Section 13.6.

## 13.2 FORECASTING FOR LONG TERM DECISIONS

### Technological Forecasting

Technological growth is often haphazard, especially in developing countries like India. This is because Technology seldom evolves and there are frequent technology transfers -due to imports of knowhow resulting in a leap-frogging phenomenon. In spite of this, it is generally seen that logarithms of many technological variables show linear trends with time, showing exponential growth. Some extrapolations reported by Rohatgi et al. are

- Passenger kms carried by Indian Airlines (Figure I)
- Fertilizer applied per hectare of cropped area (Figure II)
- Demand and supply of petroleum crude (Figure III)
- Installed capacity of electricity generation in millions of KW (figure IV).

Figure I: Passenger Km Carried by Indian Air Lines

**Figure I: Passenger Km Carried by Indian Air Lines**

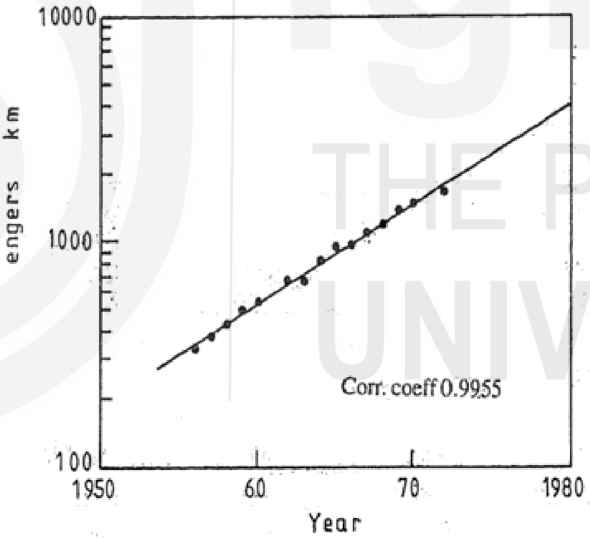
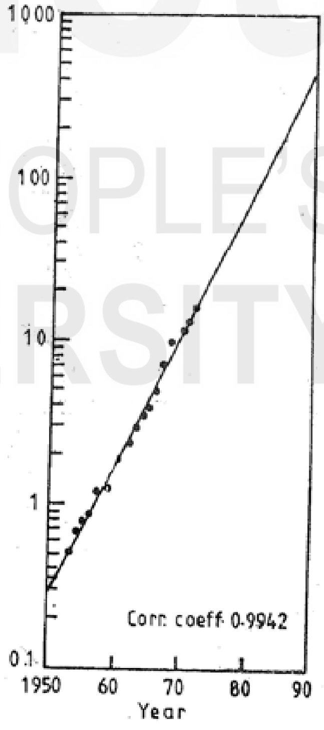


**Figure II: Fertilizer Applied per Hectare «f Cropped Area**



ignou
THE PEOPLE'S
UNIVERSITY

**Figure III: Demand and Supply of Petroleum Crude.**

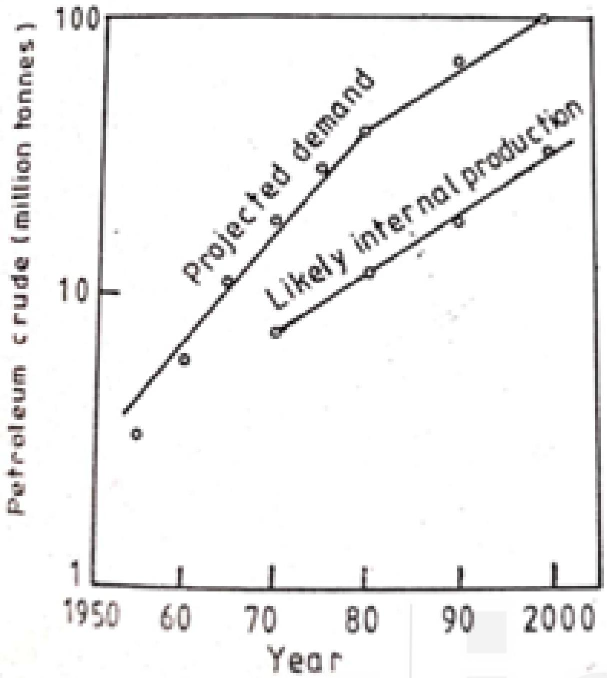
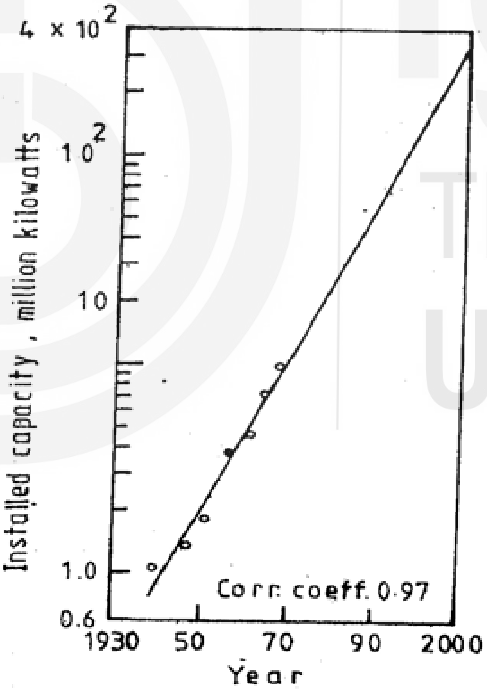


**Figure IV: Installed Capacity of Electricity Generation in Million KW**



The use of S curves in forecasting technological growth is also common. Rather than implying unchecked growth there is a limit to growth. Thus the growth rate of technology is slow to begin with (owing to initial problems), it reaches a maximum (when the technology becomes stable and popular) and finally declines till the technology becomes obsolete and is replaced by a newer alternative. Some examples of the use of S curves as reported by Rohatgi et al. (1979) are

- Hydroelectric power generation using Gumpertz growth curve (Figure V)

- Number of villages electrified using a Pearl type growth curve (Figure VI).

**Figure V: Hydroelectric Power Generalion Using Gompertz Growth Curve**

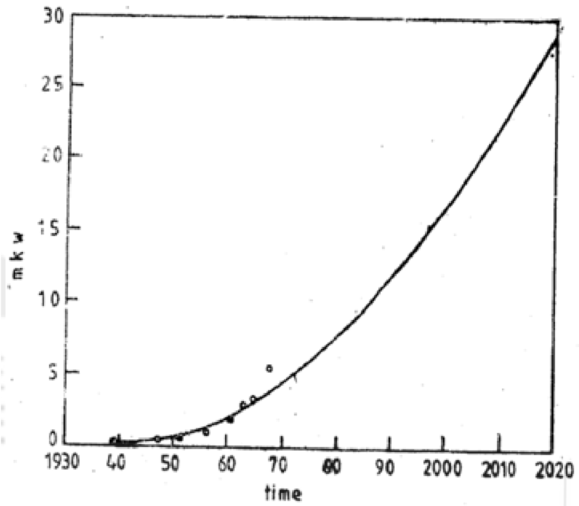
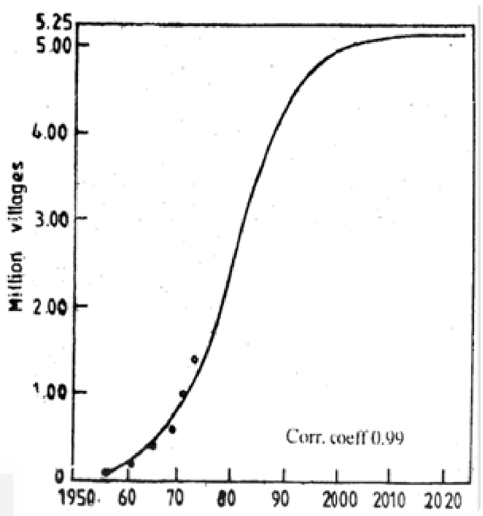


**Figure VI: Number of Villages Electrified Using a Pearl Type Growth Curve**



Apart from the above extrapolative techniques which are based on the projection of historical data into the future (such models are called regression models and you will learn more about them in Unit 15), technological forecasting often implies prediction of future scenarios or likely possible futures. As an example suppose there are three events; $E_1$, $E_2$ and $E_3$ where each one may or may not happen in the future. Thus, eight possible scenarios-$E_1$ $E_2$ $E_3$, $E_1$ $E_2$ $\overline{E}_3$, E1 $\overline{E}_2$E3, $\overline{E}_1$E2 E3, $\overline{E}_1\overline{E}_2$E3, $\overline{E}_1$E2 $\overline{E}_3$, $E_1$ $\overline{E}_2\overline{E}_3$, $\overline{E}_1$ $\overline{E}_2$ $\overline{E}_3$, — showt he range of possible futures (a line above the event indicates that the event does not take place). Moreover these events may not be independent. The breakout of war ($E_1$) is likely to lead to increased spendings on defence ($E_2$) and reduced emphasis on rural uplift and social development ($E_3$). Such interactions can be investigated using the Cross-impact Technique.

### Delphi

This is a subjective method relying on the opinion of experts designed to minimise bias and error of judgment. A Delphi panel consists of a number of experts with an impartial leader or coordinator who organises the questions.

Specific questions (rather than general opinions) with yes-no or multiple type answers or specific dates/events are sought from the experts. For instance, questions could be of the following kind :

- When do you think the petroleum reserves of the country would be exhausted? (2020,2040, 2060)
- When would the level of pollution in Delhi exceed danger limit? (as defined by a particular agency)?
- What would the population of India be in 2020, 2040 and 2060?
- When would fibre optics become a commercial viability for communication?

A summary of the responses of the participants is sent to each expert participating in the Delphi panel after a statistical analysis. For a forecast of when an event is likely to happen, the most optimistic and pessimistic estimates along with a distribution of other responses is given to the participant. On the basis of this information the experts may like to revise their earlier estimates and give revised estimates to the coordinator. It may be mentioned that the identities of the experts are not revealed to each other so that bias or influence by reputation is kept to a minimum. Also the feedback response is statistical in nature without revealing who made which forecast. The Delphi method is an iterative procedure in which revisions are carried out by the experts till the coordinator gets a stable response.

The method is very efficient, if properly conducted, as it provides a systematic framework for collecting expert opinion. By virtue of anonymity, statistical analysis and feedback of results and provision for forecast revision, results obtained are free of bias and generally reliable. Obviously, the background of the experts and their knowledge of the field is crucial. This is where the role of the coordinator in identifying the proper experts is important.

**Opinion Polls**

Opinion polls are a very common method of gaining knowledge about consumer tastes, responses to a new product, popularity of a person or leader, reactions to an election result or the likely future prime minister after the impending polls. In any opinion poll two things are of primary importance. First, the information that is sought and secondly the target population from whom the information is sought. Both these factors must be kept in mind while designing the appropriate mechanism for conducting the opinion poll. Opinion polls may be conducted through

- Personal interviews.
- Circulation of questionnaires.
- Meetings in groups.
- Conferences, seminars and symposia.

The method adopted depends largely on the population, the kind of information desired and the budget available. For instance, if information from a very large number of people is to be collected a suitably designed questionnaire could be mailed to die people concerned. Designing a proper questionnaire is itself a major task. Care should be taken to avoid ambiguous questions. Preferably, the responses should be short one word answers or ticking an appropriate reply from a set of multiple choices. This makes the questionnaire easy for the respondent to fill and also easy for the analyst to analyse. For example, the final analysis could be summarised by saying

- 80% of the population expressed opinion A,
- 10% expressed opinion B,
- 5% expressed opinion C,
- 5% expressed no opinion.

Similarly in the context of forecasting of product demand, it is common to arrive at the sales forecast by aggregating the opinion of area salesmen. The forecast could be modified based on some kind of rating for each salesman or an adjustment for environmental uncertainties.

Decisions in the area of future R&D or new technologies too are based on the opinions of experts. The Delphi method treated in this Section is just an example of a systematic gathering of opinion of experts in the concerned field.

The major advantage of opinion polls lies in the fact that a well formed opinion considers the multifarious subjective and objective factors which may not even be possible to enumerate explicitly, and yet they may have a bearing on the concerned forecast or question. Moreover the aggregation of opinion polls tends to eliminate the bias that is bound to be present in any subjective, human evaluation. In fact for long term decisions, opinion polls of opinions of the experts constitute a very reliable method for forecasting and planning.

## 13.3 FORECASTING FOR MEDIUM AND SHORT TERM DECISIONS

Forecasting for the medium and short term horizons from one to six months ahead is commonly employed for production planning, scheduling and financial planning decisions in an organisation. These methods are generally better structured as compared to the models for long term forecasting treated in Section 13.2, as the variables to be forecast are well known and often historical data is available to guide in the making of a more reliable forecast. Broadly speaking we can classify these methods into five categories.

1) Subjective of intuitive methods.
2) Methods based on an averaging of past data (moving average and exponential smoothing).
3) Regression models on historical data.
4) Causal or econometric models.
5) Stochastic models, with Time Series analysis and Box-Jenkins models.

**Subjective or Intuitive Methods**

These methods rely on the opinion of the concerned people and are quite popular in practice. Top executives, salesmen, distributors, and consumers could all be approached to give an estimate of the future demand of a product. And a judicious aggregation/adjustment of these opinions could be used to arrive at the forecast of future demand. How such opinion polls could be systematically conducted has already been discussed in Section 17.2. Committees or even a Delphi panel could be constituted for the purpose. However, all such methods suffer from individual bias and subjectivity. Moreover the underlying logic of forecast generation remains mysterious for it relies entirely on the intuitive judgment and experience of the forecaster. It cannot be documented and programmed for use on a computer so that no matter whether A or B or C makes the forecast, the result is the same. The other categories of methods discussed in the section are characterised by well laid procedures so that documentation and computerisation can be easily done.

However, subjective and intuitive methods have their own advantages. The opinion of an expert or an experienced salesman carries with it the accumulated wisdom of experience and maturity which may be difficult to incorporate in any explicit mathematical relationship developed for purposes of forecasting. Moreover in some instances where no historical data is available (e.g. forecasting the sales of a completely new product or new technology) reliance on opinions of persons in Research and Development, Marketing or other functional areas may be the only method available to forecast and plan future operations.

### Methods Based on Averaging of Past Data (Moving Averages and Exponential Smoothing)

In many instances, it may be reasonable to forecast the demand for the next period by taking the average demand till date. Similarly when the next period demand actually becomes known, it would be used in making the forecast of the next future period. However, rather than use the entire past history in determining the average. Only the recent data for the past 3 or 6 months may be used. This is the idea behind the 'Moving Average', where only the demand of the recent couple of periods (the number of periods being specified) is used in making a forecast. Consider, for illustration, the monthly sales figures of an item, shown in Table 1.

ignou THE PEOPLE'S UNIVERSITY

**Table 1: Monthly Sales of an Item and Forecasts Using Moving Averages**

| Month | Demand | 3 period moving Average | 6 period moving Average |
|---|---|---|---|
| Jan | 199 | | |
| Feb | 202 | | |
| Mar | 199 | 200 00 | |
| Apr | 208 | 203.00 | |
| May | 212 | 206 33 | |
| Jun | 194 | 203 66 | 202 33 |
| July | 214 | 205 66 | 207 83 |
| Aug | 220 | 208 33 | 210 83 |
| Sept | 2 19 | 2 16 66 | 213 13 |
| Oct | 234 | 223 33 | 217 46 |
| Nov | 219 | 223 00 | 218 63 |
| Dec | 233 | 227.66 | 225.13 |

The average of the sales for January, February and March is (199+202+199)/3=200, which constitutes the 3 months moving average calculated at the end of March and may thus be used as a forecast for April. Actual sales in April turn out to be 208 and so the 3 months moving average forecast for May is (202+199+208)/3 =203. Notice that a convenient method of updating the moving average is

$$\text{New moving average} = \text{Old moving average} + \frac{\text{Added period demand} - \text{Dropped period demand}}{\text{Number of period in moving average}}$$

At the end of May, the actual demand for May is 212, while the demand for February which is to be dropped from the last moving average is 202. Thus,

New moving average = 203 + 10/3 = 206.33 which is the forecast for June. Both the 3 period and 6 period moving average are shown in Table 1.

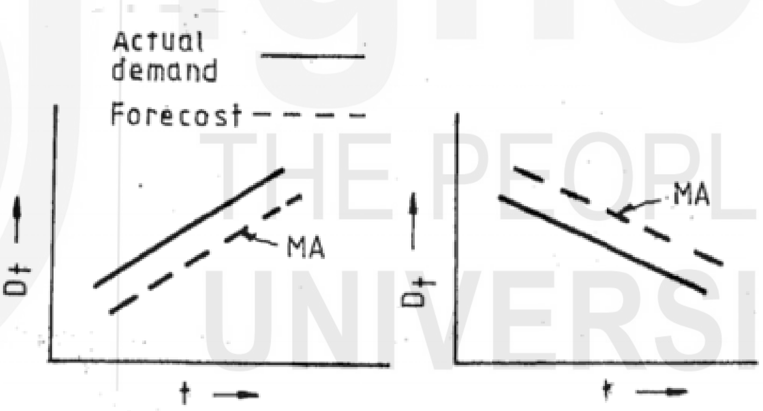
It is characteristic of moving averages to

a) Lag a trend (that is, give a lower value for an upward trend and a higher value for a lower trend) as shown in Figure VII (a).

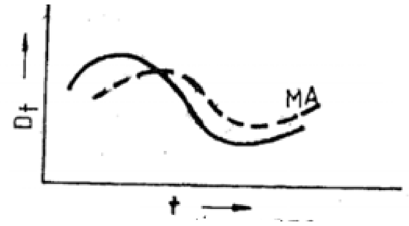
b) Be out of phase (that is, lagging) when the data is cyclic, as in seasonal demand. This is depicted in Figure VII (b).

c) Flatten the peaks of the demand pattern as shown in Figure VII (c).

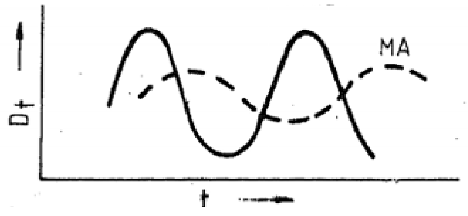
ignou
THE PEOPLE'S
UNIVERSITY



**(A) Moving Averages Lag A Trend**



**B) Moving Averages Are Out Of Phase For Cyclic Demand.**



**(C) Moving Averages Flatten Peaks**

Some correction factors to rectify the lags can be incorporated. For details, you may refer to Brown (3).

Exponential smoothing is an averaging technique where the weightage given to the past data declines (at an exponential rate) as the data recedes into the past. Thus all the values are taken into consideration, unlike in moving averages, where all data points prior to the period of the Moving Average are ignored.

If $F_t$ is the one-period ahead forecast made at time t and is the demand for period t, then

$$F_t = \mathrm{F}_{t-1} + \alpha(-D_t - F_{t-1})$$
$$= \alpha D_1 + (1-\alpha)F_{t-1}$$

Where $\alpha$ is a smoothing constant that lies between 0 and 1 but generally chosen values lie between 0.01 and 0.30. A higher value of a places more emphasis on recent data. To initiate smoothing, a starting value of $F_t$, is needed which is generally taken as the first or some average demand value available. Corrections for trend effects may be made by using double exponential smoothing and other factors. For details, you may consult the references at the end.

A computation of the smoothed values of demand for the example considered earlier in Table 1 is shown in Table 2 for values of a equal to 0.1 and 0.3. In these computations, exponential smoothing is initiated from June with a starting forecast as the average demand for the first five months. Thus the error for June is (194-204), that is -10, which when multiplied by a (0.1 or 0.3 as the case may be) and added to the previous forecast of 204 yields 203 or 201 (depending on whether $\alpha$ is 0.1 or 0.3) respectively as shown in Table 2.

**Table 2: Monthly Sales of an Item and Forecasts Using Exponential Smoothing**

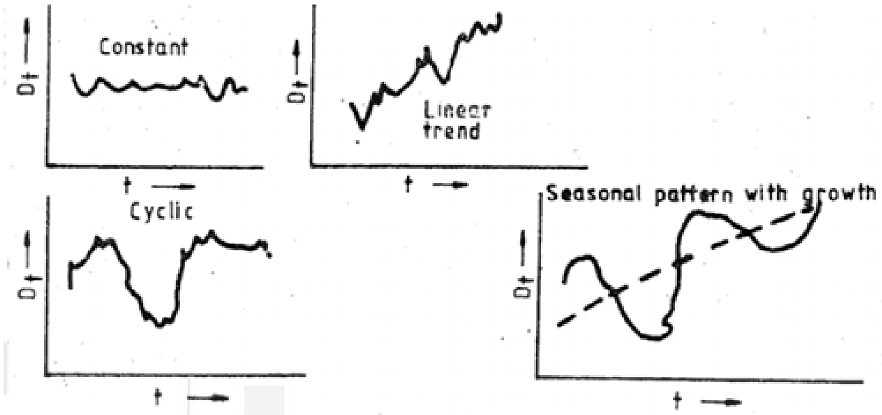
| Month | Demand | Smoothed forecast (alpha = 0.1) | Smoothed forecast (alpha = 0.3) |
|---|---|---|---|
| Jan | 199 | | |
| Feb | 202 | | |
| Mar | 199 | | |
| Apr | 208 | | |
| May | 212 | | |
| Jun | 194 | 204.0 | 204.0 |
| July | 214 | 203.0 | 201.0 |
| Aug | 220 | 204.1 | 204.9 |
| Sept | 2 19 | 205.7 | 209.4 |
| Oct | 234 | 207.0 | 212.3 |
| Nov | 219 | 209.7 | 218.8 |
| Dec | 233 | 210.6 | 218.9 |

Both moving averages and smoothing methods are essentially short term forecasting techniques where one or a few period-ahead forecasts are obtained.

### Regression Models on Historical Data

The demand of any product or service when plotted as a function of time yields a time series whose behaviour may be conceived of as following a certain pattern with random fluctuations. Some commonly observed demand patterns are shown in Figure VIII.

**Figure VIII: Some Commonly Observed Demand Patterns**



The basic approach in this method is to identify an underlying pattern and to fit a regression line to demand history by available statistical methods. The method of least squares is commonly used to determine the parameters of the fitted model.

Forecasting by this technique assumes that the underlying system of chance causes which was operating in the past would continue to operate in the future as well. The forecast would thus not be valid under abnormal conditions like wars, earthquakes, depression or other natural calamities like floods or drought which might drastically affect the variable of interest.

For the demand history considered previously in Tables 1 and 2, the linear regression line is $F_t = 193 + 3t$

where t = 1 refers to January, t=2 to February, and so on. The forecast for any month t can be found by substituting the appropriate value oft. Thus, the expected demand for next January (t=13) = 193 + (3 x 13) = 232.

You will study details of this regression procedure in Unit 19. We may only add here that the procedure can be used to fit any type of function, be it linear, parabolic or other, and that some very useful statements of confidence and precision can also be made.

### Causal or Econometric Models

In causal models, an attempt is made to consider the cause effect relationships and the variable of interest (e.g. demand) is modelled as a function of these causal variables. For instance, in trying to forecast the demand of tyres of a particular kind in a certain month (say DTM), it would be reasonable to assume that this is influenced by the targeted production of new vehicles for that month (TPVM) and the total road mileage of existing vehicles in the past 6 months (say) which could be assumed to be proportional to sales of petrol in the last 6 months (SPL6M). Thus, one possible model to forecast the monthly demand of tyres is DTM=a x (TPVM) + b x (SPL6M) where a, b and c are constants to be determined from the data. The above model has value for forecasting only if TPVM and SPL6M (the two causal variables) are known at the time the forecast is desired. This requirement is expressed by saying that these variables be leading. Also the quality of it is determined by the correlation between the predictor and the predicted variables. Commonly used indicators of the economic climate, such as consumers price index, wholesale price index, gross national product, population and per capital income are often used in econometric models because these are easily available from published records.

Model parameters are estimated by usual regression procedures, similar to the ones described in Models on Historical Data :

Construction of these structural and econometric models is generally difficult and more time-consuming as compared to simple time-series regression models. Nevertheless, they possess the advantage of portraying the inner mechanics of the demand so that when changes in a certain pertinent factor occur, the effect can be predicted.

The main difficulty in causal models is the selection or identification of proper variables which should exhibit high correlation and be leading for effective forecasting.

**Time Series Analysis or Stochastic Models**

The demand or variable of interest when plotted as a function of time yields what is commonly called a 'time-series'. This plot of demand at equal time intervals may show random patterns of behaviour and our objective in Models on Historical Data was to identify the basic underlying pattern that should be used to explain the data. After hypothesising a model (linear, parabolic or other) regression was used to estimate the model parameters, using the criterion of minimising the sum of squares of errors.

Another method often used in time series analysis is to identify the following four major components in a time series.

1) Secular trend (e.g. long term growth in market)
2) Cyclical fluctuation (e.g. due to business cycles)
3) Seasonal variation (e.g. Woollens, where demand is seasonal)
4) Random or irregular variation.

The observed value of the time series could then be expressed as a product (or some other function) of the above factors.

Another treatment that may be given to a time series is to use the framework developed by Box and Jenkins (1976) in which a stochastic model of the autoregressive (AR) variety, moving average (MA) variety, mixed autoregressive- moving average variety (ARMA) or an integrated autoregressive-moving average variety (ARIMA) model may be chosen. An introductory discussion of these models is included in Unit 20. Stochastic models are inherently complicated and require greater efforts to construct.

However, the quality of forecasting generally improves. Computer codes are available to implement the procedures [see for instance Box and Jenkins (1976)].

## 13.4 FORECAST CONTROL

Whatever, be the system of forecast generation, it is desirable to monitor the output of such a system to ensure that the discrepancy between the forecast and actual values of demand lies within some permissible range of random variations.

A system of forecast generation is shown in Figure IX.

From past data, the system generates a forecast which is subject to modification through managerial judgment and experience. The forecast is compared with the current data when it becomes available and the error is watched or monitored to assess the adequacy of the forecast generation system.

The Moving Chart is a useful statistical device to monitor and verify the accuracy of a forecasting system.

The control chart is easy to construct and maintain. Suppose data for n periods is available. If $F_{t}$,.is the forecast for period t and D, is the actual demand for period t then MR (Moving

$$\text{Range ) } =| (F_t - D_t) - (F_{t-1} - D_{t-1})|$$

∴ $\overline{\text{MR}} = \frac{\Sigma\text{MR}}{\text{n}-1}$ (There are $n - 1$ moving averages for n -periods)

Then Upper Control Limit (UCL) $= +2.66\overline{MR}$

Lower Control Limit (LCL) $= -2.66\overline{MR}$

The variable to be plotted on the chart is the error ($F_t$ – $D_t$) in each period. A sample control chart is shown in Figure X. Such a control chart tells three important things about a demand pattern:

**Figure IX: System of Forecast Generation**

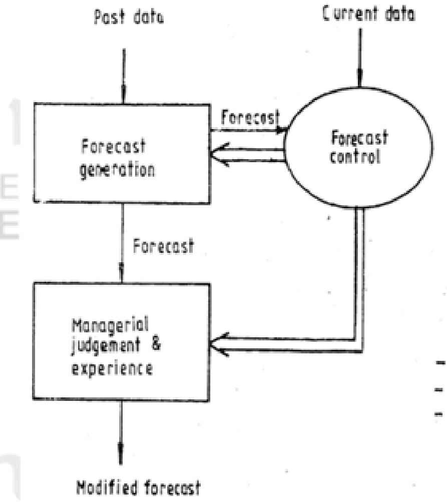
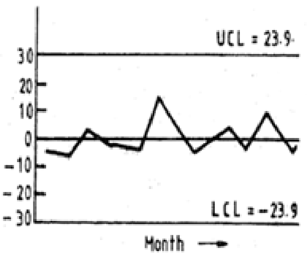


Figure X: A Control Chart Based on Moving Range



( Control chart for a sample situation

a) whether the past demand is statistically stable,

b) whether the present demand is following the past pattern,

c) if the demand pattern has changed, the control chart tells how to revise the forecasting method.

As long as the plotted error points keep falling within the control limits, it shows that the variations are due to chance causes and the underlying system of forecast generation is acceptable. When a point goes out of control there is reason to suspect the validity of the forecast generation system, which should be revised to reflect these changes.

## 13.5 SUMMARY

The unit has emphasised the importance of forecasting in all planning decisions-be they long term, medium term or short term. For long term planning decisions, techniques like Technological Forecasting, collecting opinions of experts as in Delphi or opinion polls using personal interviews or questionnaires have been surveyed. For medium and short term decisions, apart from subjective and intuitive methods there is a greater variety of mathematical models and statistical techniques that could be profitably employed. There are methods like Moving averages or exponential smoothing that are based on averaging of past data. Any suitable mathematical function or curve could be fitted to the demand history by using least squares regression. Regression is also used in estimation of parameters of causal or econometric models. Stochastic models using Box-Jenkins methodology are a statistically advanced set of tools capable of more accurate forecasting. Finally, forecast control is very necessary to check whether the forecasting system is consistent and effective. The moving range chart has been suggested for its simplicity and ease of operation in this regard.

## 13.6 SELF-ASSESSMENT EXERCISES

1) Why is forecasting so important in business? Identify applications of forecasting for

- Long term decisions.
- Medium term decisions.
- Short term decisions.

2) How would you conduct an opinion poll to determine student reading habits and preferences towards daily newspapers and weekly magazines?

3, 4, 5 For the demand data of a product, the following figures for last year's sales (monthly) are given :

ignou THE PEOPLE'S UNIVERSITY

**Period (Monthly)**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 100 | 79 | 98 | 95 | 104 | 80 | 98 | 102 | 96 | 115 | 88 |
| 67 | 53 | 601 | 79 | 102 | 118 | 135 | 162 | 70 | 53 | 68 | 63 |
| 117 | 124 | 95 | 228 | 274 | 248 | 220 | 130 | 109 | 128 | 125 | 134 |

a) Plot the data on a graph and suggest an appropriate model that could be used for forecasting.

b) Plot a 3 and 5 period moving average and show on the graph in (a)

c) Initiate exponential smoothing from the first period demand for smoothing constant (cc) values of 0.1 and 0.3. Show the plots.

6 What do you understand by forecast control? What could be the various methods to ensure that the forecasting system is appropriate?

## 13.7 KEY WORDS

**Causal Models:** Forecasting models wherein the demand or variable or interest is related to underlying causes or causal variables.

**Delphi:** A method of collecting information from experts, useful for long term forecasting. It is iterative in nature and maintains anonymity to reduce subjective bias.

**Exponential Smoothing:** A short term forecasting method based on weighted averages of past data so that the weightage declines exponentially as the data recedes into the past, with the highest weightage being given to the most recent data.

**Forecasting:** A systematic procedure to determine the future value of a variable of interest.

**Moving Average:** An average computed by considering the K most recent (for a K- period moving average) demand points, commonly used for short term forecasting.

**Prediction:** A term to denote the estimate or guess of a future variable that may be arrived at by subjective hunches or intuition.

**Regression:** From a given demand history to establish a relation between the dependent variable (such as demand) and independent variable (S). Such relations prove very useful for forecasting purposes.

**Time Series:** Any data on demand, sales or consumption taken at regular intervals of time constitutes a time series. Analysis of this time series to discover patterns of growth, decay, seasonalities or random fluctuations is known as to Time Series analysis.

ignou
THE PEOPLE'S
UNIVERSITY

## 13.8 FURTHER READINGS

Biegel, J.E., *Production Control-A Quantitative Approach*, Prentice Hall of India: New Delhi.

Box, G.E.P. and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*, I-lolden-Day: San Francisco.

Brown, R.G., *Smoothing, Forecasting and Prediction of Discrete Time Series*, Prentice Hall: Englewood-Cliffs.

Chambers, J.C., S.K. Mullick and D.D. Smith. An *Executive's Guide to Forecasting*, John Wiley: New York.

Chatterjee, S., & Simonoff, J.S. Handbook of regression analysis (Vol. 5). John Wiley & Sons.

Firth, M., *Forecasting Methods in Business and Management*, Edward Arnold: London.

Jarrett, Al., *Forecasting for Business Decisions*, Basil Blackwell: London.

Makridakis, S. and S. Wheelwright. *Forecasting: Methods and Applications*, John Wiley: New York.

Martino. J.P., *Technological Forecasting for Decision Making*, American Elsevier: New York, .

Montgomery D.C. and L.A. Johnson, *Forecasting and Time Series Analysis*, McGraw Hill: New York.

Rohatgi. P.K., K. Rohatgi and B. Bowonder. *Technological Forecasting*, Tata McGraw Hill: New Delhi.

# UNIT 14 CORRELATION

## Objectives

After completion of this unit, you should be able to :

- understand the meaning of correlation
- compute the correlation coefficient between two variables from sample observations
- test for the significance of the correlation coefficient
- identify confidence limits for the population correlation coefficient from the observed sample correlation coefficient
- compute the rank correlation coefficient when rankings rather than actual values for variables are known
- appreciate some practical applications of correlation
- become aware of the concept of auto-correlation and its application in time series analysis.

## Structure

14.1 Introduction

14.2 The Correlation Coefficient

14.3 Testing for the Significance of the Correlation Coefficient

14.4 Rank Correlation

14.5 Practical Applications of Correlation

14.6 Auto-correlation and Time Series Analysis

14.7 Summary

14.8 Self-assessment Exercises

14.9 Key Words

14.10 Further Readings

## 14.1 INTRODUCTION

We often encounter situations where data appears as pairs of figures relating to two variables. A correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter. The regression problem, which is treated in Unit 15, considers the frequency distributions of one variable (called the dependent variable) when another (independent variable) is held fixed at each of several levels.

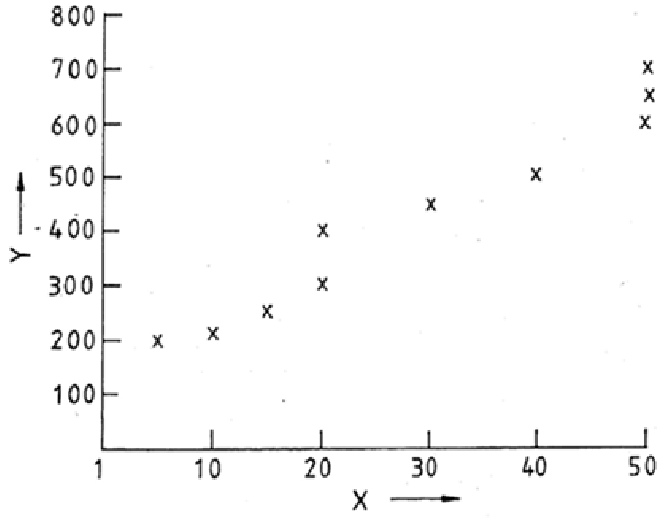
Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage marks obtained by a person in SSC examination, blood pressure and metabolism or the relation between height and weight of individuals. In these examples both variables are observed as they naturally occur, since neither variable is fixed at predetermined levels.

Examples of regression problems can be found in the study of the yields of crops grown with different amount of fertiliser, the length of life of certain animals exposed to different amounts of radiation, the hardness of plastics which are heat-treated for different periods of time, and so on. In these problems the variation in one measurement is studied for particular levels of the other variable selected by the experimenter. Thus the factors or independent variables in regression analysis are not assumed to be random variables, though the dependent variable is modelled as a random variable for which intervals of given precision and confidence are often worked out. In correlation analysis, all variables are assumed to be random variables. For example, we may have figures on advertisement expenditure (X) and Sales (Y) of a firm for the last ten years, as shown in Table I. When this data is plotted on a graph as in Figure I we obtain a scatter diagram. A **scatter diagram** gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of what kind of relationship (linear or non-linear) would describe the relationship. Correlation examines the first question of determining whether an association exists between the two variables, and if it does, to what extent. Regression examines the second question of establishing an appropriate relation between the variables.

**Table 1: Yearwise data on Advertisement Expenditure and Sales**

| Year | Advertisement Expenditure in Thousand Rs. (X) | Sales in thousand Rs. (Y) |
|---|---|---|
| 1988 | 50 | 700 |
| 1987 | 50 | 650 |
| 1986 | 50 | 600 |
| 1985 | 40 | 500 |
| 1984 | 30 | 450 |
| 1983 | 20 | 400 |
| 1982 | 20 | 300 |
| 1981 | 15 | 250 |
| 1980 | 10 | 210 |
| 1979 | 5 | 200 |

**Figure I: Scatter Diagram**



The scatter diagram may exhibit different kinds of patterns. Some typical patterns indicating different correlations between two variables are shown in Figure II.

What we shall study next is a precise and quantitative measure of the degree of association between two variables and the correlation coefficient.

## 14.2 THE CORRELATION COEFFICIENT

**Definition and Interpretation**

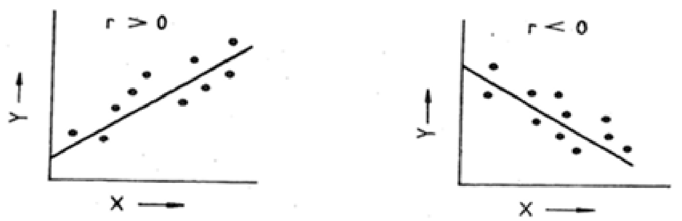
The correlation coefficient measures the degree of association between two variables X and Y. Pearson's formula for correlation coefficient is given as

$$r = \frac{\frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})}{\sigma_x \sigma_y} \qquad (14.1)$$

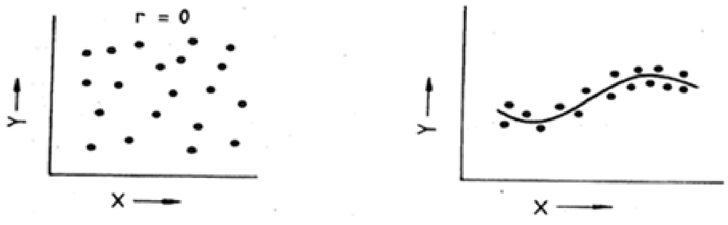
Where r is the correlation coefficient between X and Y, $\sigma_x$ and $\sigma_y$ are the standard deviations of X and Y respectively and n is the number of values of the pair of variable X and Y in the given data. The expression $\frac{1}{n}\Sigma(X-(\bar{X})(Y-\bar{Y})$ is known as the **covariance** between X and Y. Here r is also called the Pearson's product moment correlation coefficient. You should note that r is a dimensionless number whose numerical value lies between +1 and -1. Positive values of r indicate positive (or direct) correlation between the two variables X and Y i.e. as X increases Y will also increase or as X decreases Y will also decrease. Negative values of r indicate negative (or inverse) correlation, thereby meaning that an increase in one variable results in a decrease in the value of the other variable. A zero correlation means that there is no association between the two variables. Figure II shows a number of scatter plots with corresponding values for the correlation coefficient r.

ignou
THE PEOPLE'S
UNIVERSITY

**Figure II: Different Types of Association Between Variables**



(a) Positive Correlation (b) Negative Correlation



(c) No Correlation (d) Non-linear association

The following form for carrying out computations of the correlation coefficient is perhaps more convenient

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \qquad (14.2)$$

where

$x = X - \bar{X}$ = deviation of a particular X value from the mean $\bar{X}$

$y = Y - \bar{Y}$ = deviation of a particular Y value from the mean $\bar{Y}$

Equation (14.2) can be derived from equation (14.1) by substituting for $\sigma_x$ and $\sigma_y$ as follows:

$$\sigma_n = \sqrt{\frac{1}{n}\Sigma(X-\bar{X})^2} \text{ and } \sigma_y = \sqrt{\frac{1}{n}\Sigma(X-\bar{Y})^2} \qquad (14.3)$$

**Activity A**

Suggest five pairs of variables which you expect to be positively correlated.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

**Activity B**

Suggest five pairs of variables which you expect to be negatively correlated.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

ignou THE PEOPLE'S UNIVERSITY

A **Sample Calculation**: Taking as an illustration the data of advertisement expenditure (X) and Sales (Y) of a company for the 10-year period shown in Table 1, we proceed to determine the correlation coefficient between these variables:

Computations are conveniently carried out as shown in Table 2.

**Table 2: Calculation of Correlation Coefficient**

| Sl.No . | X | Y | $x = \mathrm{x} - \bar{\mathrm{x}}$ | $y = y - \bar{y}$ | $\mathrm{x}^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|---|
| 1. | 50 | 700 | 21 | 274 | 441 | 75,076 | 5,7 |
| 2. | 50 | 650 | 21 | 224 | 441 | 50,176 | 4,7 |
| 3. | 50 | 600 | 21 | 174 | 441 | 30,276 | 3,6 |
| 4. | 40 | 500 | 11 | 74 | 121 | 5,476 | 814 |
| 5. | 30 | 450 | 1 | 24 | 1 | 576 | 24 |
| 6. | 20 | 400 | -9 | -26 | 81 | 676 | 234 |
| 7. | 20 | 300 | -9 | -126 | 81 | 15,876 | 1,134 |
| 8. | 15 | 250 | -14 | -176 | 196 | 30,976 | 2,464 |
| 9. | 10 | 210 | -19 | -216 | 361 | 46,656 | 4,104 |
| 10. | 5 | 200 | -24 | -226 | 576 | 51,076 | 5,424 |
| Total | 290 | 4,260 | 0 | 0 | 2,740 | 3,06,840 | 28,310 |

$$\bar{X} = \frac{290}{10} = 29$$

$$\bar{\mathrm{Y}} = \frac{4260}{10} = 426$$

$$\therefore r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{28310}{\sqrt{2740 \times 306840}} = 0.976$$

This value of r (= 0.976) indicates a high degree of association between the variables X and Y. For this particular problem, it indicates that an increase in advertisement expenditure is likely to yield higher sales.

You may have noticed that in carrying out calculations for the correlation coefficient in Table 2, large values for $x^2$ and $y^2$ resulted in a great computational burden. Simplification in computations can be adopted by calculating the deviations of the observations from an assumed average rather than the, actual average, and also scaling these deviations conveniently. To illustrate this short cut procedure, let us compute the correlation coefficient for the same data. We shall take U to be the deviation of X values from the assumed mean of 30 divided by 5. Similarly, V represents the deviation of Y values from the assumed mean of 400 divided by 10.

The computations are shown in Table 3.

ignou
THE PEOPLE'S
UNIVERSITY

**Table 3: Short cut Procedure for Calculation of Correlation Coefficient**

| S.No | X | Y | U | V | UV | $U^2$ | $V^2$ |
|---|---|---|---|---|---|---|---|
| 1. | 50 | 700 | 4 | 30 | 120 | 16 | 900 |
| 2. | 50 | 650 | 4 | 25 | 100 | 16 | 625 |
| 3 | 50 | 600 | 4 | 20 | 80 | 16 | 400 |
| 4. | 40 | 500 | 2 | 10 | 20 | 4 | 100 |
| 5. | 30 | 450 | 0 | 5 | 0 | 0 | 25 |
| 6. | 20 | 400 | -2 | 0 | 0 | 4 | 0 |
| 7 | 20 | 300 | -2 | -10 | 20 | 4 | 100 |
| 8. | 15 | 250 | -3 | -15 | 45 | 9 | 225 |
| 9. | 10 | 210 | -4 | -19 | 76 | 16 | 361 |
| 10. | 5 | 200 | -5 | -20 | 100 | 25 | 400 |
| Total | | | -2 | 26 | 561 | 110 | 3,13 |

$$r = \frac{\Sigma UV - \frac{\Sigma U \Sigma V}{n}}{\sqrt{\Sigma U^2 - \frac{(\Sigma U)^2}{n}}\sqrt{\Sigma V^2 - \frac{(\Sigma V)^2}{n}}}$$

$$r = \frac{561 - \frac{(-2)(26)}{10}}{\sqrt{110 - \frac{(-2)^2}{10}}\sqrt{3136 - \frac{(26)^2}{10}}}$$

$$= \frac{566.2}{10.47 \times 55.39}$$

$$= 0.976$$

We thus obtain the same result as before.

**Activity C**

Use the short cut procedure to obtain the value of correlation coefficient in the above example using scaling factor 10 and 100 for X and Y respectively. (That is, the deviation from the assumed mean is to be divided by 10 for X values and by 100 for Y values.)

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

## 14.3 TESTING FOR THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

Once the correlation coefficient has been calculated from sample data one is normally interested in asking the question: Is there an association between

ignou THE PEOPLE'S UNIVERSITY

the variables? Or with what confidence can we make a statement about the association between the variables?

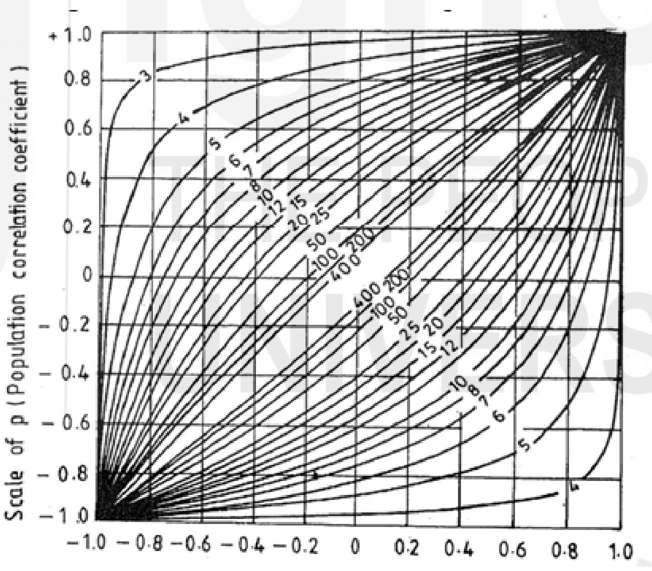
Such questions are best answered statistically by using one of the following two commonly used procedures :

1) Providing confidence limits for the population correlation coefficient from the sample size n and the sample correlation coefficient r. If this confidence interval includes the value zero, then we say that r is not significant, implying thereby that the population correlation coefficient may be zero and the value of r may be due to sampling variability.
2) Testing the null hypothesis that population correlation coefficient equals zero vs. the alternative hypothesis that it does not, by using the t-statistic.

The use of both these procedures is now illustrated.

The value of the sample correlation coefficient is used as an estimate of the true population correlation p. It is desirable to incude a confidence interval for the true value along with the sample statistics. There are several methods for obtaining the confidence interval for p. However, the most straight forward method is to use a chart such as that shown in Figure III.

**Figure III: Confidence Bands for the Population Correlation**



**Scale of r ( Sample correlation cofficient)**

Once r has been calculated, the chart can be used to determine the upper and lower values of the interval for the sample size used. In this chart the range of unknown values of p is shown in the vertical scale; while the sample r values are shown on the horizontal axis, with a number of curves for selected sample sizes. Notice that for every sample size there are two curves. To read the 95% confidence limits for an observed sample correlation coefficient of 0.8 for a sample of size 10, we simply look along the horizontal line for a value of 0.8 (the sample correlation coefficient) and construct a vertical line from there till it intersects the first curve for n =10. This happens for p = 0.2. This is the lower limit of the confidence interval. Extending the vertical line upwards, it again intersects the second n =10 line at p = 0.92, which represents the upper confidence limit. Thus the 95% confidence interval for the population correlation coefficient becomes

$$0.2 \leq p \leq 0.92$$

If a confidence interval for p includes the value zero, then r is not considered significant since that value of r may be due to nothing more than sampling variability.

This method of using charts to determine the confidence intervals is convenient, though of course we must use a different chart for different confidence limits (e.g. 90%, 95%, 99%).

The alternative approach for testing the significance of r is to use the formula

$$t = \frac{r}{\sqrt{(1-r^2)(n-2)}} \qquad (14.4)$$

Referring to the table of t-distribution for (n-2) degrees of freedom, we can find the critical value for t at any desired level of significance (5% level of significance is commonly used). If the calculated value of t (as obtained by equation 14.3) is less than or equal to the table value, we accept the hypothesis ($H_0$: the correlation coefficient equals zero), meaning that the correlation between the variables is not significantly different from zero:

Suppose we obtain a correlation coefficient of 0.2 for a sample of size 10.

$$t = \frac{0.2}{\sqrt{(1-0.04)/8}} \cong 0.577$$

And from the t-distribution with 8 degrees of freedom for a 5% level of significance, the table value = 2.306. Thus we conclude that this r of 0.2 for n = 10 is not significantly different from zero.

It should be mentioned here that in case the same value of the correlation coefficient of 0.2 was obtained on a sample of size 100 then

$$\mathrm{t} = \frac{0.2}{\sqrt{(1-0.04)/98}} \cong 2021$$

And the tabled value for a t-distribution with 98 degrees of freedom and a 5% level of significance = 1.99. Since the calculated t exceeds this figure of 1.99, we can conclude that this correlation coefficient of 0.2 on a sample of size 100 could be considered significantly different from zero, or alternatively that there is statistically significant association between the variables.

## 14.4 RANK CORRELATION

Quite often data is available in the form of some ranking for different variables. It is common to resort to rankings on a preferential basis in areas such as food testing, competitive events (e.g. games, fashion shows, or beauty contests) and attitudinal surveys. The primary purpose of computing a correlation coefficient in such situations is to determine the extent to which the two sets of rankings are in agreement. The coefficient that is determined from these ranks is known as Spearman's rank correlation coefficient, r.

This is given by the following formula

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{14.5}$$

Here n is the number of pairs of observations and $d_i$ is the difference in ranks for the i$^{th}$ observation set.

Suppose the ranks obtained by a set of ten students in a Mathematics test (variable X) and a Physics test (variable Y) are as shown below :

| Rank for variable X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank for variable Y | 3 | 1 | 4 | 2 | 6 | 9 | 8 | 10 | 5 | 7 |

To determine the rank correlation, $r_s$ we can organise computations as shown in Table 4 :

**Table 4: Determination of Spearman's Rank Correlation**

| Individual | Rank in Maths(X) | Rank in Physics(Y) | d =Y -X | $d^2$ |
|---|---|---|---|---|
| 1 | 1 | 3 | +2 | 4 |
| 2 | 2 | 1 | -I | 1 |
| 3 | 3 | 4 | +1 | 1 |
| 4 | 4 | 2 | -2 | 4 |
| 5 | 5 | 6 | +1 | 1 |
| 6 | 6 | 9 | +3 | 9 |
| 7 | 7 | 8 | +1 | 1 |
| 8 | 8 | 10 | +2 | 4 |
| 9 | 9 | 5 | -4 | 16 |
| 10 | 10 | 7 | -3 | 9 |
| Total | | | | 50 |

Using the formula (14.5) we obtain

$$r_s = 1 - \frac{6 \times 50}{10(100-1)} = 1 - 0.303 = 0.697$$

We can thus say that there is a high degree of correlation between the performance in Mathematics and Physics.

We can also test the significance of the value obtained. The null hypothesis is that the two variables are not associated, i.e. r = O. That is, we are interested to test the null hypothesis, $H_o$ that the two variables are not associated in the population and that the observed value of $r_s$ differs from zero only by chance. The t-statistic that is used to test this is

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

$$= 0.697\sqrt{\frac{10-2}{1-(0.697)^2}}$$

=2.75

Referring to the table of the t-distribution for n-2 = 8 degrees of freedom, the critical value for t at a 5% level of significance is 2.306. Since the calculated value of t is higher than the table value, we reject the null hypothesis concluding that the performances in Mathematics and Physics are closely associated.

When two or more items have the same rank, a correction has to be applied to $\sum \mathrm{d}_{\mathrm{i}}^2$. For example, if the ranks of X are 1, 2, 3, 3, 5, ... showing that there are two items with the same 3rd rank, then instead of writing 3, we write $3\frac{1}{2}$ for each so that the sum of these items is 7 and the mean of the ranks is unaffected. But in such cases the standard deviation is affected, and therefore, a correction is required. For this, $\sum \mathrm{d}_{\mathrm{i}}^2$ is increased by $(t^3 - t)/12$ for each tie, where t is the number of items in each tie.

**Activity D**

Suppose the ranks in Table 4 were tied as follows: Individuals 3 and 4 both ranked 3rd in Maths and individuals 6, 7 and 8 ranked 8th in Physics. Assuming that other rankings remain unaltered, compute the value of Spearman's rank correlation.

………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

………………………………………………………………………………………

## 14.5 PRACTICAL APPLICATIONS OF CORRELATION

The primary purpose of correlation is to establish an association between any two random variables. The presence of association does not imply causation, but the existence of causation certainly implies association. Statistical evidence can only establish the presence or absence of association between variables. Whether causation exists or not depends merely on reasoning. For example, there is reason to believe that higher income causes higher expenditure on superior quality cloth. However, one must be on the guard against spurious or nonsense correlation that may be observed between totally unrelated variables purely by chance.

Correlation analysis is used as a starting point for selecting useful independent variables for regression analysis. For instance a construction company could identify factors like

- population
- construction employment
- building permits issued last year which it feels would affect its sales for the current year.

These and other factors that may be identified could be checked for mutual correlation by computing the correlation coefficient of each pair of variables from the given historical data (this kind of analysis is easily done by using an appropriate routine on a computer). Only variables having a high correlation with the yearly sales could be singled out for inclusion in a regression model.

Correlation is also used in factor analysis wherein attempts are made to resolve a large set of measured variables in terms of relatively few new Categories, known as factors. The results could be useful in the following three ways :

1) to reveal the underlying or latent factors that determine the relationship between the observed data,
2) to make evident relationships between data that had been obscured before such analysis, and
3) to provide a classification scheme when data scored on various rating scales have to be grouped together.

Another major application of correlation is in forecasting with the help of time series models. In using past data (which is often a time series of the variable of interest available at equal time intervals) one has to identify the trend, seasonality and random pattern in the data before an appropriate forecasting model can be built. The notion of auto-correlation and plots of auto-correlation for various time lags help one to identify the nature of the underlying process. Details of time series analysis are discussed in Unit 20. However, some fundamental concepts of auto-correlation and its use for time series analysis-are outlined below.

## 14.6 AUTO-CORRELATION AND-TIME SERIES ANALYSIS

The concept of auto-correlation is similar to that of correlation but applies to values of the same variable at different time lags. Figure IV shows how a single variable such as income (X) can be used to construct another variable ($X_1$) whose only difference from the first is that its values are lagging by one time period. Then, X and $X_1$ can be treated as two variables and their correlation found. Such a correlation is referred to as auto-correlation and shows how a variable relates to itself for a specified time lag. Similarly, one can construct $X_2$ and find its correlation with X. This correlation will indicate how values of the same variable that are two periods apart relate to each other.

**Figure IV: Example of the Same Variable with Different Time Lags**

| Time | X Original variable | X1One time lag variable constructed from X | X2 Two time lags variable constructed from X |
|---|---|---|---|
| t = 1 | 13 | | |
| 2 | 8 | 8 | |
| 3 | 15 | 15 | 15 |
| 4 | 4 | 4 | 4 |
| 5 | 4 | 4 | 4 |
| 6 | 12 | 12 | 12 |
| 7 | 11 | 11 | 11 |
| 8 | 7 | 7 | 7 |
| 9 | 14 | 14 | 14 |
| 10 | 12 | 12 | 12 |

One could construct from one variable another time-lagged variable which is twelve periods removed. If the data consists of monthly figures, a twelve-month time lag will show how values of 'the same month but of different years correlate with each other. If the auto-correlation coefficient is positive, it implies that there is a seasonal pattern of twelve months duration. On the other hand, a near zero auto-correlation indicates the absence of a seasonal pattern. Similarly, if there is a trend in the data, values next to each other will relate, in the sense that if one increases, the other too will tend to increase in order to maintain the trend. Finally, in case of completely random data, all auto-correlations will tend to zero (or not significantly different from zero).

The formula for the auto correlation coefficient at time lag k is:

$$\mathrm{r_k} = \frac{\sum_{t=1}^{n-k} (\mathrm{X_t} - \overline{\mathrm{X}})(\mathrm{X_{t+k}} - \overline{\mathrm{X}})}{\sum_{t=1}^{n} (\mathrm{X_t} - \overline{\mathrm{X}})^2}$$

where

$r_k$ denotes the auto-correlation coefficient for time lag k k denotes the length of the time lag n is the number of observations

X, is the value of the variable at time t and

X is the mean of all the data

Using the data of Figure IV the calculations can be illustrated.

$$\bar{X} = \frac{13 + 8 + 15 + \cdots + 12}{10} = \frac{100}{10} = 10$$

$$r_1 = \frac{(13-10)(8-10) + (8-10)(15-10) + \cdots + (14-10)(12-10)}{(13-10)^2 + (8-10)^2 + \cdots + (14-10)^2 + (12-10)^2}$$

$$= \frac{-27}{144} = -0.188$$

For k = 2, the calculation is as follows :

$$r_2 = \frac{\sum_{t=1}^{8\overline{7}} (X_t - 10)(X_{t-2} - 10)}{\sum_{t=1}^{10} (X_t - 10)^2}$$

$$= \frac{(13-10)(15-10) + (8-10)(4-10) + \cdots + (7-10)(12-10)}{(13-10)^2 + (8-10)^2 + \cdots + (14-10)^2 + (12-10)^2}$$

$$= \frac{-29}{144} = -.201$$

A plot of the auto-correlations for various lags is often made to identify the nature of the underlying time series. We, however, reserve the detailed discussion on such plots and their use for time series analysis for Unit 16.

## 14.7 SUMMARY

In this unit the concept of correlation or the association between two variables has been discussed. A scatter plot of the variables may suggest that the two variables are related but the value of the Pearson correlation coefficient r quantifies this association. The correlation coefficient r may assume values between -1 and 1. The sign indicates whether the association is direct (+ve) or inverse (-ve). A numerical value of r equal to unity indicates perfect association while a value of zero indicates no association.

Tests for significance of the correlation coefficient have been described. Spearman's rank correlation for data with ranks is outlined. Applications of correlation in identifying relevant variables for regression, factor analysis and in forecasting using time series have been highlighted. Finally the concept of auto-correlation is defined and illustrated for use in time series analysis.

## 14.8 SELF-ASSESSMENT EXERCISES

1) What do you understand by the term correlation? Explain how the study of correlation helps in forecasting demand of a product.

2) A company wants to study the relation between R&D expenditure (X) and annual profit (Y). The following table presents the information for the last eight years:

| Year | R&D Expense (X) (Rs. in thousands) | Annual Profit (Y) |
|---|---|---|
| 1988 | 9 | 45 |
| 1987 | 7 | 42 |
| 1986 | 5 | 41 |
| 1985 | 10 | 60 |
| 1984 | 4 | 30 |
| 1983 | 5 | 34 |
| 1982 | 3 | 25 |
| 1981 |  | 20 |

a) Plot the data on a scatter diagram.

b) Estimate the sample correlation coefficient.

c) What are the 95% confidence limits for the population correlation coefficient?

d) Test the significance of the correlation coefficient using a t-test at a significance level of 5%.

3) The following data pertains to length of service (in years) and. the annual income for a sample of ten employees of an industry:

| **Length of service in years (X)** | **Annual income in thousand rupees (Y)** |
|---|---|
| 6 | 14 |
| 8 | 17 |
| 9 | 15 |
| 10 | 18 |
| 11 | 16 |
| 12 | 22 |
| 14 | 26 |
| 16 | 25 |
| 18 | 30 |
| 20 | 34 |

Compute the correlation coefficient between X and Y and test its significance at levels of 0.01 and 0.05.

4) Twelve salesmen are ranked for efficiency and the length of service as below:

| **Salesman** | **Efficiency (X)** | **Length of Service (Y)** |
|---|---|---|
| A | 1 | 2 |
| B | 2 | 1 |
| C | 3 | 5 |
| D | 5 | 3 |
| E | 5 | 9 |
| F | 5 | 7 |
| G | 7 | 7 |
| H | 8 | 6 |
| I | 9 | 4 |
| J | 10 | 11 |
| K | 11 | 10 |
| L | 12 | 11 |

a) Find the value of Spearman's rank correlation coefficient, $r_s$

b) Test for the Significance of $r_s$

5) An alternative definition of the correlation coefficient between a two-dimensional random variable (X, Y) is

$$\rho = \frac{[(X - E(X))(Y - E(Y))]}{\sqrt{V(X)V(Y)}}$$

where E(.) represents expectation and V(.) the variance of the random variable. Show that the above expression can be simplified as follows :

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

(Notice here that the numerator is called the covariance of X and Y).

6) In studying the relationship between the index of industrial production and index of security prices the following data from the Economic Survey 1980-81 (Government of India Publication) was collected.

| | 70-71 | 71-72 | 72-73 | 73-74 | 74-75 | 75-76 | 76-77 | 77-78 | 78-79 |
|---|---|---|---|---|---|---|---|---|---|
| Index of Industrial (1970-100) | 101.3 | 114.8 | 1196.6 | 122.1 | 125.2 | 122.2 | 135.3 | 140.1 | 150.1 |
| Index of Security Prices (1970-71-100) | 100.0 | 95.1 | 96.7 | 116.0 | 113.2 | 96.9 | 102.9 | 107.4 | 130.4 |

a) Find the correlation between the two indices.

b) Test the significance of correlation coefficient at 0.01 level of significance.

7) Compute and plot the first five auto-correlations (i.e. up-to time lag 5 periods) for the time series given below :

| t | I | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| dt | 13 | 8 | 15 | 4 | 4 | 12 | II | 7 | 14 | 12 |

## 14.9 KEY WORDS

**Auto-correlation:** Similar to correlation in that it described the association or mutual dependence between values of the same variable but at different time periods. Auto-correlation coefficients provide important information about the structure of a data set.

**Correlation:** Degree of association between two variables.

**Correlation Coefficient :** A number lying between -1 (Perfect negative correlation) and + 1 (perfect positive correlation) to quantify the association between two variables.

**Covariance:** This is the joint variation between the variables X and Y. Mathematically defined as

ignou THE PEOPLE'S UNIVERSITY

$$\frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

for n data points.

**Scatter Diagram:** An ungrouped plot of two variables, on the X and Y axes.

**Time Lag:** The length between two time periods, generally used in time series where one may test, for instance, how values of periods 1, 2; 3, 4 correlate with values of periods 4, 5, 6, 7 (time lag 3 periods).

**Time-Series:** Set of observations at equal time intervals which may form the basis of future forecasting.

## 14.10 FURTHER READINGS

Box, G.E.P., and G.M. Jenkins. *Time Series Analysis, Forecasting and Control*, Holden-Day: San Francisco.

Chatterjee, S., & Simonoff, J.S. Handbook of regression analysis (Vol.5). John Wiley & Sons.

Draper, N. and H. Smith. *Applied Regression Analysis*, John Wiley: New

York.

Edwards, B., *The Readable Maths and Statistics Book,* George Allen and Unwin: London.

Makridakis, S. and S. Wheelwright. *Interactive Forecasting: Univariate and Multivariate Methods*, Holden-Day: San Francisco.

Peters, W.S. and G.W: Summers. *Statistical Analysis for Business Decisions*, Prentice Hall: Englewood-Cliffs.

Srivastava, U.K., G.V. Shenoy and S.C. Sharma. *Quantitative Techniques for Managerial Decision Making,Wiley* Eastern: New Delhi.

Stevenson, W.J. *Business Statistics-Concepts and Applications*, Harper and Row: New York.

ignou THE PEOPLE'S UNIVERSITY

# UNIT 15 REGRESSION

**Objectives**

After successful completion of this unit, you should be able to:

- understand the role of regression in establishing mathematical relationships between dependent and independent variables from given data
- use the least squares criterion to estimate the model parameters
- determine the standard errors of estimate of the forecast and estimated parameters
- establish confidence intervals for the forecast values and estimates of parameters
- make meaningful forecasts from given data by fitting any function, linear in unknown parameters.

**Structure**

15.1 Introduction
15.2 Fitting A Straight Line
15.3 Examining the Fitted Straight Line
15.4 An Example of the Calculations
15.5 Variety of Regression Models
15.6 Summary
15.7 Self-assessment Exercises
15.8 Key Words
15.9 Further Readings

ignou THE PEOPLE'S UNIVERSITY

## 15.1 INTRODUCTION

In industry and business today, large amounts of data are continuously being generated. This may be data pertaining, for instance, to a company's annual production, annual sales, capacity utilisation, turnover, profits, manpower levels, absenteeism or some other variable of direct interest to management. Or there might be technical data regarding a process such as temperature or pressure at certain crucial points, concentration of a certain chemical in the product or the breaking strength of the sample produced or one of a large number of quality attributes.

The accumulated data may be used to gain information about the system (as for instance what happens to the output of the plant when temperature is reduced by half) or to visually depict the past pattern of behaviour (as often happens in company s annual meetings where records of company progress are projected) or simply used for control purposes to check if the process or system is operating as designed (as for instance in quality control). Our interest in regression is primarily for the first purpose, mainly to extract the main features of the relationships hidden in or implied by the mass of data.

## The Need for Statistical Analysis

For the system under study there may be many variables and it is of interest to examine the effects that some variables exert (or appear to exert) on others. The exact functional relationship between variables may be too complex but we may wish to approximate to this functional relationship by some simple mathematical function such as straight line or a polynomial which approximates to the true function over certain limited ranges of the variables involved.

There could be many variables of interest in the system. In a chemical plant for instance, the monthly consumption of water or other raw materials, the temperature and pressure maintained in the reacting vessel, the number of operating days per month the monthly production of the final product and any by-products could all be variables of interest. We are, however, interested in some key performance variable (which in our case may be monthly production of final product) and would like to see how this key variable (called the response variable or dependent variable) is affected by the other variables (often called independent variables). By independent variables we shall usually mean variables that can either be set to a desired value or else take values that can be observed but not controlled. As a result of changes that are deliberately made, or simply take place in the independent variables, an effect is transmitted to the response variables. In general we shall be interested in finding out how changes in the independent variables affect the values of the response variables. Sometimes the distinction between independent and dependent variables is not clear, but a choice may be made depending on convenience or objectives.

Broadly speaking we would have to undergo the following sequence of steps in determining the relationship between variables, assuming we have data points already.

1) Identify the independent and response variables.
2) Make a guess of the form of the relation (linear, quadratic, cyclic etc.) between the dependent and independent variables. This can be facilitated by a graphical plot of the data (for two variables) on a systematic tabulation (for more than two variables) which may suggest some trends or patterns.
3) Estimate the parameters of the tentatively entertained model in step 2 above. For instance if a straight line was to be fitted, what is the slope and intercept of this line?
4) Having obtained the mathematical model, conduct an error analysis to see how good the model fits into the actual data.
5) Stop, if satisfied with model otherwise repeat steps 2 to 4 for another choice of the model form in step 2.

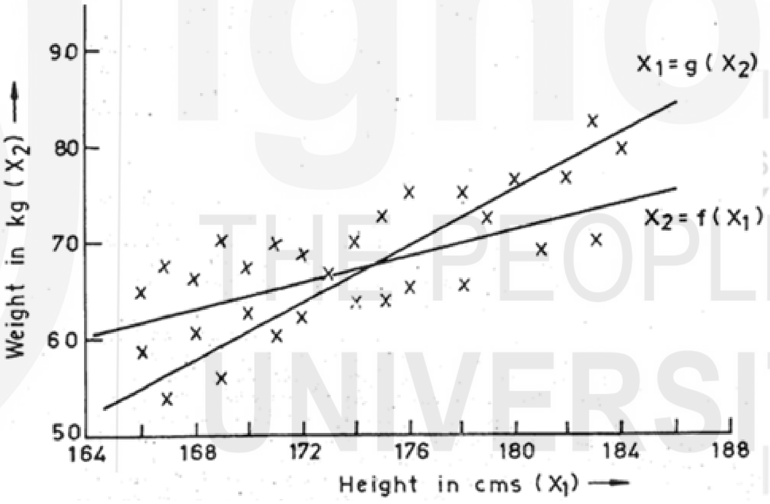
ignou THE PEOPLE'S UNIVERSITY

### What is Regression?

Suppose we consider, the height and weight of adult males for some given population. If we plot the pair $(X_1, X_2)$ = (height, weight), a diagram like Figure I will result. Such a diagram, you would recall from the previous chapter, is conventionally called a scatter diagram.

Note that for any given height there is a range of observed weights and vice-versa. This variation will be partially due to measurement errors but primarily due to variations between individuals. Thus no unique relationship between actual height and weight can be expected. But we can note that average observed weight for a given observed height increases as height increases. The locus of average observed weight for given observed height (as height ' varies) is called the regression curve of weight on height. Let us denote it by $X_2 = f(X_1)$. There also exists a **regression curve** of height on weight similarly defined which we can denote by $X_I = g(X_2)$. Let us assume that these two "curves" are both straight lines (which in general they may not be). In general these two curves are not the same as indicated by the two lines in Figure I.

**Figure I: Height and Weight of Thirty Adult Males**



A pair of random variables such as (height, weight) follows some sort of bivariate probability distribution. When we are concerned with the dependence of a random variable Y on quantity X, which is variable but not a random variable, an equation that relates Y to X is usually called a **regression equation**. Similarly when more than one independent variable is involved, we may wish to examine the way in which a response Y depends on variables $X_1 X_2 \ldots . X$. We determine a regression equation from data which cover certain areas of the X-space as $Y = f(X_1, X_2 \ldots X_k)$

### Linear Regression

The simplest and most commonly used relationship between two variables is that of a straight line. We may write the linear, first order model as

$$Y = \beta_o + \beta_1 X + \epsilon \qquad (15.1)$$

That is, for a given X, a corresponding observation Y consists of the value $\beta_0 + \beta_1 X$ plus an amount $\in$, the increment by which an individual Y may fall off the regression line. Equation (15.1) is the model of what we believe $\beta_0, \beta_1$ are called the parameters of the model whose values have been obtained from the actual data.

When we say that a model is linear or non-linear, we are referring to linearity or non¬linearity in the parameters. The value of the highest power of independent variable in the model is called the order of the model. For example:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

is a second order (in X) linear (in the his) regression model..

Now in the model of equation (15.1) $\beta_0, \beta_1$ and ∈ are unknown and in fact ∈ would be difficult to discover since it changes from observation to observation. However, $\beta_0$ and $\beta_1$ remain fixed, and although we cannot find them exactly without examining all possible occurrences of Y and X, we can use the information provided by the actual data to give us estimates $b_o$ and $b_1$ of $\beta_0$ and $\beta_1$. Thus we can write

$$Y = b_o + b_1 X \qquad (15.2)$$

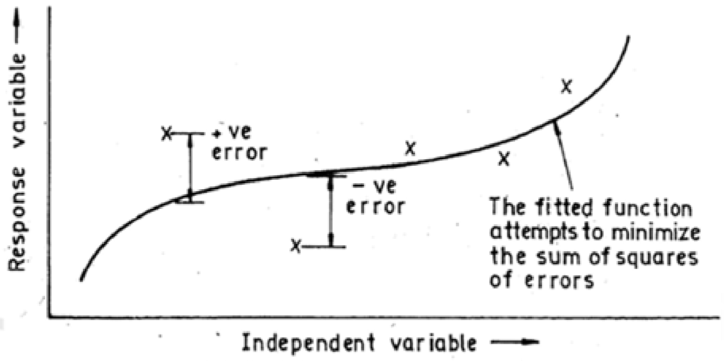
where Y that denotes the predicted value of Y for a given X, when $b_o$ and $b_1$ are determined. Equation 15.2 could then be used as a predictive equation; substitution of a value of X would provide a. prediction of the true mean value of Y for that X.

## 15.2 FITTING A STRAIGHT LINE

### Least Squares Criterion

In fitting a straight line (or any other function) to a set of data points we would expect some points to fall above or below the line resulting in both positive and negative error terms (see Figure II). It is true that we would like the overall error to be as small as possible. The most common criterion in the determination of model parameters is to minimise the sum of squares of errors, or residuals as they are often called. This is known as the least squares criterion, and is the one most commonly used in regression analysis.

**Figure II: The Least Squares Criterion**


Response variable
+ ve error
– ve error
The fitted function attempts to minimize the sum of squares of errors
Independent variable

This is, however, not the only criterion available. One may, for instance, minimise the sum of absolute deviations, which is equivalent to minimising the mean absolute deviation (MAD). The least squares criterion, however, has the following main advantages :

1) It is simple and intuitively appealing.
2) It results in linear equations (called normal equations) for solution of parameters which are easy to solve.
3) It results in estimates of quality of fit and intervals of confidence of predicted values rather easily.

In the context of the straight line model of equation (15.1), suppose there are n data points $(X_1Y_1), (X_2Y_2), \dots, (X_n, Y_n)$ then we can write from equation (19.1)

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i, i = 1, \dots n \tag{15.3}$$

so that the sum of squares of the deviations from the true line is

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{15.4}$$

We shall choose our estimates $b_0$ and $b_1$ to be values which, when substituted for $\beta_0$ and $\beta_1$ in equation (15.4) produce the least possible value of S. We can determine $b_0$ and $b_1$ by differentiating equation (15.4) first with respect to $\beta_o$ and then with respect to $\beta_1$ and setting the results equal to zero. Notice that $X_i$, $Y_i$ are fixed pairs of numbers from our data set for i varying between 1 and n. Therefore,

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(Y_{ij} - \beta_o - \beta_1 X_i\right)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

so that the estimates $b_o$ and $b_1$ are given by

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X) = 0$$

$$\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X) = 0$$

where we substitute $(b_o, b_1)$ for $(\beta_0, \beta_1)$ when we derivatives to zero.

We thus obtain two linear equations in two unknown parameters $(\beta_0, \beta_1)$. These equations are known as normal equations and for this case they can be written as

$$\begin{Bmatrix} b_0 n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{Bmatrix} \tag{15.5}$$

The solution to these equations is easily written as follows:

$$b_0 = \frac{\begin{vmatrix} \Sigma Y_i & \Sigma X_i \\ \Sigma X_i Y_i & \Sigma X_i^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{vmatrix}} = \frac{\sum Y_i \Sigma X_i^2 - \Sigma X_i \Sigma X_i Y_i}{n \Sigma X_i^2 - (\Sigma X_i)^2} \qquad (15.6)$$

$$b_1 = \frac{\begin{vmatrix} n & \Sigma Y_i \\ \sum X_i & \Sigma X_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \Sigma X_i \\ \Sigma x_i & \Sigma x_i^2 \end{vmatrix}} = \frac{n \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n \Sigma X_i^2 - (\Sigma X_i)^2} \qquad (15.7)$$

Thus (15.6) and (15.7) may be used to determine the estimates of the parameters and the predictive equation (15.2) may be used to obtain the predicted value of Y (called Y) for any desired value of X.

Rather than use the above procedure, a slightly modified (though equivalent) method is to use the, solution of the first normal equation in (15.5) to obtain boas

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad (15.8)$$

Where $\bar{X}$ and $\bar{Y}$ are $(X_1 + X_2 + \cdots + X_n)/n$ and $(Y_1 + Y_2 + \cdots + Y_n)/n$ respectively. Substituting (15.8) in (15.2) yields the following estimated regression equation

$$\hat{Y} = \bar{Y} + b_1 (X - \bar{X}) \qquad (15.9)$$

Where $b_1$ is computed by

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2} \qquad (15.10)$$

This equation, as you can easily see, is derived from the last expression in (15.7) by simply dividing the numerator and denominator by n. It is written in the form above as it has an interpretation suitable for analysis of variance later.

### Activity A

You can see that the last form of equation (15.10) is expressed in terms of sums of squares or products of deviations of individual points from their corresponding means. Show that in fact

$$\text{fact } \Sigma(\mathrm{X_i} - \bar{\mathrm{X}})(\mathrm{Y_i} - \bar{\mathrm{Y}}) = \Sigma \mathrm{X_i Y_i} - (\Sigma \mathrm{X_i Y_i})/\mathrm{n}$$

and $\Sigma (X_i - \bar{X})^2 = \Sigma X_i^2 - (\Sigma \mathrm{X})^2/n$

Hence verify equation (15.10).

The quantity $\mathrm{X_i^2}$ is called the uncorrected sum of squares of the $\mathrm{X'}$ s, and $(\sum X_i)^2/n$ is the correction for the mean of the $\mathrm{X'}$ s. The difference is called the corrected sum of squares of the $\mathrm{X'}$ s. Similarly, $\sum \mathrm{x_i Y_i}$ is called uncorrected sum of products, and $(\sum X_i \sum Y_i)/n$ is the correction for the means of X and Y. The difference is called the corrected sum of products of X and Y. In terms of these definitions we can see that the estimate of the slope of the fitted Straight line, bi from equation 15.10, is simply the ratio of the corrected sum of products of X and Y to the corrected sum of squares of X's.

**How good is the Regression?**

**Analysis of Variance (ANOVA)** Once the regression line is obtained we would like to find out how good tie fit is. This can be ascertained by the examination of errors. If $Y_i$ is the ith data point and $\hat{Y}_i$ its predicted value by the regression equation, then we can write

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - \left(\hat{Y}_i - \bar{Y}\right)$$

If we square both sides and add the equations for i = 1 to n, we obtain

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left\{(Y_i - \bar{Y}) - \left(\hat{Y}_i - \bar{Y}\right)\right\}^2$$

$$= \Sigma(Y_i - \bar{Y})^2 + \Sigma\left(\left(\hat{Y}_i - \bar{Y}\right)^2 - 2\Sigma(Y - \bar{Y})\left(\hat{Y}_i - \bar{Y}\right)\right.$$

The third term can be rewritten as

$$= -2\Sigma(Y_i - \bar{Y})\left(\hat{Y}_i - \bar{Y}\right) \qquad \text{(using eqn. 15.9)}$$

$$= -2\Sigma(Y_i - \bar{Y})b_1(X_i - \bar{X}) \qquad \hat{Y} - \bar{Y} = b_1(X - \bar{X})$$

$$= -2b_1\Sigma(Y_i - \bar{Y})(X_i - \bar{X}) \qquad \text{(using eqn. 15.10)}$$

$$= -2b_1^2\Sigma(x_i - \bar{x})^2$$

$$= -2\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$$

Thus

$$\Sigma\left(Y_i - \hat{Y}\right)^2 = \Sigma(Y_i - \bar{Y}_i)^2 + \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2 \qquad (15.11)$$

which may be written as

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma\left(Y_i - \hat{Y}_i\right)^2 + \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2 \text{ (using eqn. 15.11)}$$

Now $Y_i - \hat{Y}$ is the deviation of the ith observation from the overall mean and so the left hand side of equation (15.11) is the sum of squares of the deviations of the observations from the mean; this is shortened to SS (SS: Sum of squares) about the mean, and is also the corrected sum of squares of the Y's. Since $Y_i - \hat{Y}_i$ is the deviation of the ith observation from its predicted or fitted value, and $\hat{Y}_i - \bar{Y}_\imath$ is the deviation of the predicted value of the ith observation from the mean, we can express equation (15.11) in words as follows :

$$\begin{pmatrix}\text{Sum of squares}\\ \text{about the mean}\end{pmatrix} = \begin{pmatrix}\text{Sum of squares}\\ \text{about regression}\end{pmatrix} + \begin{pmatrix}\text{Sum of squares}\\ \text{due to regression}\end{pmatrix}$$

This shows that, of the variation in the Y's about their mean, some of the variation can be ascribed to the regression line and some $\sum\left(Y_i - \hat{Y}_i\right)$ to the fact that the actual observations do not all lie on the regression line. If they all did, the sum of squares about the regression would be zero. From this procedure, we can see that a way of assessing how useful the regression line will be as a predict or is to see how much of the SS about the mean has fallen into the SS about regression. We shall be pleased if the SS due to regression is much greater than the SS about regression, or what amounts to the same thing if the ratio

$$\mathrm{R}^2 = \frac{\text{SS due to regression}}{\text{SS about mean}}$$

is not too far from unity:

Any sum of, squares has associated with it a number called its degrees of freedom. This number indicates how many independent pieces of information involving the n independent numbers $\mathrm{Y}_1, \mathrm{Y}_2 \dots, \mathrm{Y}_\mathrm{n}$, are needed to compile the sum of squares. For example, the SS about the mean needs (n-1) independent pieces (for the numbers $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots \dots, Y_n - \bar{Y}$ only (n-1) are independent since all the n numbers sum to zero, by definition of the mean). We can compute the SS due to regression from a single function of $Y_1, \quad Y_2 \quad \dots \quad Y_{n,}$, namely $b_1$ (since $\sum(Y_i - \bar{Y})^2 = b_1^2 \sum(X_i - \bar{X})^2$ and so this sum of squares-has one degree of freedom.

By subtraction, the SS about regression has (n-2) degrees of freedom. Thus, corresponding to equation (15.11), we can show the split of degrees of freedom as (n - 1) = (n - 2) + 1 … (15.12)

Using equations (15.11) and (15.12) and employing alternative computational forms for the expression of equation (15.11) we can construct an analysis of variance (ANOVA) table in the following form :

**ANOVA TABLE**

| Source | Sum of Squares | Degrees of Freedom | Mean Square |
|---|---|---|---|
| Regression | $b_1\left[\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}\right]$<br>$= \frac{[\Sigma X_i Y_i - (\Sigma X_i \Sigma Y_i)/n]^2}{\Sigma X_i^2 - (\Sigma X_1)^2/n}$ | 1 | $\mathrm{MS}_\mathrm{k}$ |
| About regression (residual) | By subtraction | n-2 | $\mathrm{s}^2 = \frac{(\mathrm{SS})}{\mathrm{n}-2}$ |
| About mean (total, corrected for mean) | | ‘n - 1 | |

The Mean Square column is obtained by dividing each sum of squares-entry by its corresponding degrees of freedom. The mean square about regression, $s^2$ will provide an estimate, based on (n - 2) degrees of freedom, of the variance about the regression, a quantity we shall call $\sigma^2{}_{\mathrm{YX}}$ If the regression equation were estimated from an indefinitely large number of observations, the variance about regression would represent a measure of the error with which any observed value of Y could be predicted from a given value of X using the determined equation.

**An Example:** Data on the annual sales of a company in lakhs of Rupees over the past eleven years is shown to the Table below. Determine a suitable straight line regression model, $\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon$ for the data in the table.

| Year | Annual Sales in lakhs of Rupees |
|---|---|
| 1998 | 1 |
| 1999 | 5 |
| 2000 | 4 |
| 2001 | 7 |
| 2002 | 10 |
| 2003 | 8 |
| 2004 | 9 |
| 2005 | 13 |
| 2006 | 14 |
| 2007 | 13 |
| 2008 | 18 |

**Solution:** The independent variable in this problem is the year whereas the response variable is the annual sales. Although we could take the actual year as the independent variable itself, a judicious choice of the origin at the middle year of 2003 with the corresponding X values for other years as -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5 would simplify calculations. From equation. (15,10) we see that to estimate the parameter bj we require the four summations $\sum \mathrm{X_i}, \sum \mathrm{Y_i}, \sum \mathrm{X_i^2}$ and $\sum \mathrm{X_i Y_i}$.

Thus, calculations can be organised as shown below where the totals of the four columns yield the four desired summations:

| Year | Annual Sales in lakhs of Rupees |
|---|---|
| 1998 | 1 |
| 1999 | 5 |
| 2000 | 4 |
| 2001 | 7 |
| 2002 | 10 |
| 2003 | 8 |
| 2004 | 9 |
| 2005 | 13 |
| 2006 | 14 |
| 2007 | 13 |
| 2008 | 18 |

We find that

n = 11

$$\sum X_i = 0$$

$$\overline{X} = 0/11 = 0$$

$$\Sigma Y_{\rm i} = 102$$

$$\bar{Y} = 102/11$$

$$\Sigma X_1^2 = 110$$

$\Sigma X_i Y_i = 158$

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i \Sigma Y_i)/n}{\sum X_1^2 - (\sum X_i)^2/n}$$

$$= \frac{158}{110} = 1.44$$

The fitted equation is thus n

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

or $\hat{Y} = 9.27 + 1.44X$

Thus the parameters $\beta_0$ and $\beta_1$ of the model $\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon$ are estimated by $\mathrm{b_o}$ and $\mathrm{b_1}$ which in this case are 9.27 and 1.44 respectively. Now that the model is completely specified we can obtain the predicted values $\hat{Y}_i$ and the errors or residuals $\mathrm{Y_i} - \widehat{\mathrm{Y}}_\mathrm{i}$ corresponding to the eleven observations. These are shown in the table below:

| I | $\mathbf{X}_i$ | $\mathbf{Y}_1$ | $\hat{Y}_i$ | $Y_i - \hat{Y}_i$ |
|---|---|---|---|---|
| 1 | -5 | 1 | 2.07 | -1.07 |
| 2 | -4 | 5 | 3.51 | 1,49 |
| 3 | -3 | 4 | 4.95 | -0.95 |
| 4 | -2 | 7 | 6.39 | 0.61 |
| 5 | -1 | 10 | 7.83 | 2.17 |
| 6 | 0 | 8 | 9.27 | -1.27 |
| 7 | 1 | 9 | 10.71 | -1-71 |
| 8 | 2 | 13 | 12.15 | 0.85 |
| 9 | 3 | 14 | 13,59 | 0.41 |
| 10 | 4 | 13 | 15.03 | -2.03 |
| 11 | 5 | 18 | 16.47 | 1.53 |

To determine whether the fit is good enough, the ANOVA table can be constructed.

SS duo to regression = $b_i[\Sigma X_i Y_i - (\Sigma X_i \Sigma Y_i)/n]$

(Associated degrees of feedom = 1) = $\frac{[\sum X_{\rm i} Y_{\rm i} - (\sum X_i \Sigma Y_{\rm i})/n]^2}{\sum X_{\rm i}^2 - (\Sigma X_{\rm i})^2/n}$

$$= \frac{(158)^2}{110} = 226.95$$

The total (corrected) SS = $\Sigma Y_i^2 - (\Sigma Y_i)^2/n$
(Associated degrees of $= 1194 - (102)^2/11$
Freedom = 11 –1 = 10) $= 1194 - 945.82$
= 248.18

The value $\mathrm{R}^2 = \frac{\text{SS due to regression}}{\text{SS about mean}} = \frac{226.95}{248.18} = 0.9145$

indicating that the regression line explains 91.45% of the total variation about the mean.

**A NOVA TABLE**

| Source | SS | Df | MS |
|---|---|---|---|
| Regression (b) | 226.95 | 1 | MSR = 226.95 |
| Residual | 21.23 | 9 | $s^2 = 2.36$ |
| Total (corrected) | 248.18 | 10 | |

## 15.3 EXAMINING THE FITTEID STRAIGHT LINE

In fitting the linear model $\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon$ using the least squares criterion as indicated above in Section 15.2, no assumption were made about probability distributions. The method of estimating the parameters $\beta_0$ and $\beta_1$ tried only to minimise the sum of squares of the errors or residuals, and that simply involved the solution of simultaneous linear equations. However, in order to be able to evaluate the precision of the estimated parameters and provide confidence intervals for forecasted values, it is necessary to make the following assumptions in the basic model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1,2, \dots \dots .., n$

1) $\epsilon_1$ is a random variable with mean zero and variance $\sigma^2$ (unknown), that is $E(\epsilon_i) = 0, V(\epsilon_1) = \sigma^2$
2) $\epsilon_i$ and $\epsilon_j$ are uncorrelated, i ≠ j, so that Cov $(\epsilon_i, \epsilon_j) = 0$

   Thus $E(Y_i) = \beta_0 + \beta_1 X_i, V(Y_i) = \sigma^2$ and $Y_i$ and $\mathrm{Y_j}, \mathrm{i} \neq \mathrm{j}$, are uncorrelated. A further assumption, which is not immediately necessary and will be recalled when used, is that
3) $\epsilon_i$ is a normally distributed random variable, with mean zero and variance $\sigma^2$ by assumption (1), that is

$$\epsilon_{\mathrm{i}} \sim \mathrm{N}(0, \sigma^2)$$

Under this additional assumption $(\epsilon_i, \epsilon_j)$ are not only uncorrelated but necessarily independent.

It may be mentioned here that errors that occur in many real life situations tend to be normally distributed due to the Central Limit Theorem. In practice an error term such as ∈ is a sum of errors from several sources. Then no matter what the probability distribution of the separate errors may be, their sum will have a distribution that will tend more and more to-the normal distribution as the number of components increases, by the Central Limit Theorem. Using the above assumptions, we can determine the following :

1) Standard error of the slope b, and confidence interval for $\beta_1$,
2) Standard error of the intercept $b_o$ and a confidence interval for $\beta_0$
3) Standard error or $\hat{Y}$, the predicted value
4) Significance of regression
5) Percentage variation explained

**Standard Error of the Slope and Confidence Interval for its Estimate**

From equation (15.10)

$$b_i = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

(since the other term removed from the numerator is $\sum_{i=1}^{n} (X_i - \bar{X})\bar{Y} = \bar{Y}\sum_{i=1}^{n} (X_i - \bar{X}) = 0$

$$= [(X_i - \bar{X})Y_1 + \cdots \ldots + (X_n - \bar{X})Y_n] / \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Now the variance of a function

$$F = a_1Y_1 + a_2Y_2 + \cdots \ldots + a_nY_n \text{ is}$$

$$V(F) = a_1^2V(Y_1) + a_2^2V(Y_2) + \cdots \ldots a_n^2V(Y_n)$$

if the $Y_i$ are pairwise uncorrelated and the $a_1$ are constants, furthermore,

if $\mathrm{V(Y_i)} = \sigma^2$,

$$V(F) = (a_1^2 + a_2^2 + \cdots \ldots a_n^2)\sigma^2$$

In the expression for $a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

since the $\mathrm{X_i}$ can be regarded as constants.

Hence after reduction

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad (15.13)$$

The standard error (s.e.) of $b_1$ is the square root of the variance, that is

$$\text{s.e. } (b_1) = \frac{\sigma}{[\sum_{i=1}^{n} (X_i - \bar{X})^2]^{\frac{1}{2}}} \qquad (15.14)$$

If $\sigma$ is unknown, we may use the estimate s in its place and obtain the estimated standard error of $b_1$, as

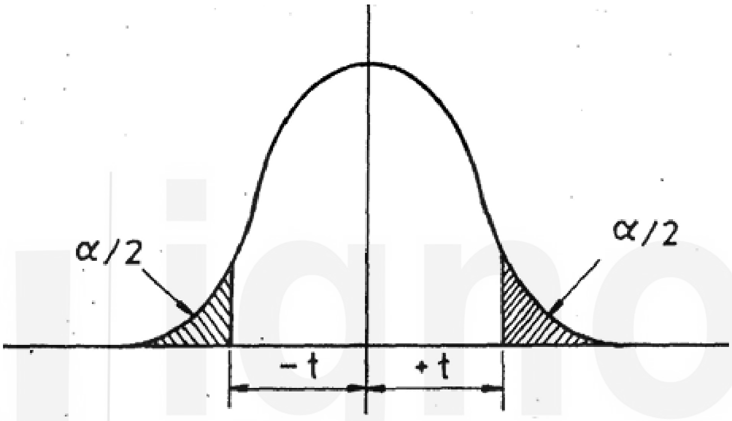
$$\text{est s.e. } (b_1) = \frac{s}{[\Sigma(X_i - \bar{X})^2]^{\frac{1}{2}}} \qquad (15.15)$$

If we assume that the variations of the observations about the line are normal, that is, that the errors e, are all from the same normal distribution, $\mathrm{N}(0,\sigma^2)$, it can be shown that we can assign $100(1-\alpha)\%$ confidence limits for $\beta_1$, by calculating

$$b_1 \pm \frac{t\left(n-2,1-\frac{1}{2}\alpha\right)s}{[(X_i-\bar{X})^2]^{\frac{1}{2}}} \ldots (15.16)$$

where $t\left(n-2,1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$ percentage point-of .a t-distribution with n -2 degrees of freedom (the number of degrees of freedom on which the estimate $s^2$ is based) (see Figure I1I)

**Figure III: The t-distribution**



**Standard Error of the Intercept and Confidence Interval for its Estimate**

We may recall from equation (15.8) that

$$b_0 = \bar{Y} - b_1\bar{X}$$

In computing the variance of bQ we require the variance of (which is $\frac{1}{n}\sum_{i=1}^{n} Y_i$ and thus has variance $\frac{1}{n^2}\sum_{i=1}^{n}$ Var $(Y_i) = \frac{\sigma^2}{n}$, since Var $(Y_i) = \sigma^2$) by assumption (2) stated at the beginning of Section (15.3) and the variance of $b_1$ (which is available from equation (15.13) above. Since X may be treated as a constant we may write.

$$\mathrm{V}(\mathrm{b}_0) = \mathrm{V}(\bar{\mathrm{Y}}) + (\bar{\mathrm{X}})^2\mathrm{V}(\mathrm{b}_1)$$

Substituting for V(Y) and $\mathrm{V}(\mathrm{b}_1)$ as indicated above, we obtain:

$$V(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i-\bar{X})^2}\right]$$

$$= \frac{\sigma^2\Sigma X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} \qquad (15.17)$$

In like manner if $\sigma^2$ is unknown, $s^2$ may be used to determine the estimated variance and standard error of $\mathrm{b_o}$ (square root of the variance). Thus the 100 (1- a)% confidence limits for $\beta_0$ are given by

$$b_0 \pm t\left(n-2,1-\frac{1}{2}\alpha\right)\left\{\frac{\Sigma x_i^2}{n\sum_{i=1}^{n}(\bar{X}_i-\bar{X})^2}\right\}^{\frac{1}{2}} s \qquad (15.18)$$

where, as before $\left(n-2,1-\frac{1}{2}\alpha\right)$ corresponds to the $1-\frac{\alpha}{2}$ percentage point of a t-distribution with (n-2) degrees of freedom(see Figure III once again)

**Standard Error of the Forecast**

The forecast or predicted value of the dependent variable Y can be expressed in terms of averages, by using equation (19.9), as

$$\hat{Y}=\bar{Y}+b_1(X-\bar{X})$$

where both Y and $b_1$ are subject to error which will influence $\hat{Y}$. Now if $\mathbf{a}_i$. and $\mathbf{C}_i$ are constants, and

$$a=a_1Y_1+a_2Y_2+\cdots+a_nY_n$$

$$C=C_1Y_1+C_2Y_2+\cdots C_nY_n$$

then provided that $Y_1$ and $Y_j$ are uncorrelated when $i\neq j$ and if $V(Y_i)=\sigma^2$, all i, Cov (a, c) = $(a_1c+a_2c_2+\cdots+a_nC_n)$

if follows by setting $a=\bar{Y}\left(i.ea_i=\frac{1}{n}\right)$ and $c=b_1($ i.e.$c_i=(X_i-\bar{X})/\sum_{i=1}^{n}(X_i-\bar{X})^2$, so that cov $\left(\bar{\hat{Y}},b_1\right)=0$, that is $\bar{Y}$ and $b_1$ are uncorrelated random variables. Thus the variance of the predicted mean value of Y, $\hat{Y}_k$ at a specific value $X_k$ of X is

$$V(\hat{Y}_k)=V(\bar{Y})+(X_k-\bar{X})^2V(b_1)$$

$$\frac{\sigma^2}{n}+\frac{(X_k-\bar{X})^2\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \qquad (15.19)$$

where the expression in equation (15.13) for V $(b_1)$ has been utilised.

Hence the estimated standard error for the predicted mean value of Y for a given $X_k$ is

$$\text{est. s.e. }(\hat{Y})=s\times\left[\frac{1}{n}+\frac{(X_k-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{\frac{1}{2}}\ \ldots (15.20)$$

This is a minimum when $X_k=\bar{X}$ and increases as we move $X_k$ away from $\bar{X}$ in either direction. In other words, the greater distance an $X_k$ is (in either direction) from X, the larger is the error we may expect to make when predicting from the regression line the mean value of Y at $X_k$ (that is $\hat{Y}_k$). This is intuitively meaningful since we expect the best predictions in the middle of our observed range of X, with predictions becoming worse as we move away from the range of observed X values.

The variance and standard error in equations (15.19) and (15.20) above apply to the predicted mean value of Y for a given $X_k$. Since the actual observed value of Y varies about the true mean value with variance $\sigma^2$ (independent of the $V(\hat{Y})$, a predicted value of an individual observation will still be given by Y but will have a variance

$$\sigma^2+V(\hat{Y}_k)=\sigma^2\left[1+\frac{1}{n}+\frac{(X_k-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]\ \ldots (15.21)$$

If $\sigma^2$ is unknown the corresponding value may be obtained by inserting $s^2$ for $\sigma^2$. In a similar fashion, the $100(1-\alpha)\%$ confidence limits for a new observation which will be centered on $Y_k$ is

$$\hat{Y}_k \pm t\left(n-2,1-\frac{\alpha}{2}\right)\left[1+\frac{1}{n}+\frac{(X_k-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{\frac{1}{3}} \ldots (15.22)$$

Where $t\left(n-2,1-\frac{\alpha}{2}\right.$ corresponds to the $\left(1-\frac{\alpha}{2}\right)$ percentage point of a t-distribution with (n-2) degrees of freedom (recall Figure III).

**F-test for Significance of Regression**

Since the Y, are random variables, any function of them is also a random variable; two particular functions are $MS_R$ the mean square due to regression, and $s^2$, the mean square due to residual variation, which arise in the analysis of variance table shown in Section 15.2.

In the case of fitting a straight line, it can be shown that if $\beta_1 = 0$ (i.e. the slope of the fitted line is zero) the variable $MS_R$ multiplied by its degree of freedom (here one), and divided by $\sigma^2$ follows $s^2$ (chi-square) distribution with the same (1) number of degrees of freedom. In addition$(n-2)s^2/\sigma^2$ follows a $\chi^2$ distribution with (n - 2) degrees of freedom. And since these two variables are independent, a statistical theorem tells us that the ratio.

$$F = \frac{MS_R}{s^2} \ldots (15.23)$$

follows an F distribution with 1 and (n - 2) degrees of freedom provided $\beta_1 = 0)$. This fact can thus be used as a test of $\beta_1 = 0$. We compare the ratio $F = MS_R/s^2$ with the 100 (1- $\alpha$)% point of the tabulated F(1, n - 2) distribution in order to determine whether $\beta_1$ can be considered non-zero on the basis of the observed data.

**Percentage Variation Explained**

The quantity $R^2$ defined earlier in Section 15.2 as the ratio of the SS due to regression to SS about the mean measures the "proportion of total variation about the mean Y explained by the regression". It is often expressed as a percentage by multiplying it by 100.

## 15.4 AN EXAMPLE OF THE CALCULATIONS

The various computations outlined in the case of a straight line regression situation in Section 15.3 will now be illustrated for the example of annual sales data for a company that was considered earlier in Section 15.2. Recall that the fitted regression equation was

$\hat{Y} = 9.27 + 1.44$ X.

By choosing any value for X the corresponding prediction $\hat{Y}$ could be made by using this equation. However, the parameters of this model have been estimated from the given data under certain assumptions, and these estimates may be subject to error. Consequently the forecast obtained is subject to chance errors. It is now our objective to

ignou THE PEOPLE'S UNIVERSITY

1) Quantify the errors of estimates of the parameters $b_o$ and b,
2) Establish reasonable confidence intervals for the parameter values n
3) Quantify the error-of the forecast $\hat{Y}_k$ Yk made at some point $X_k$
4) Provide confidence intervals for the forecasted values at some $X_k$
5) Test for the significance of regression, and
6) To obtain an overall measure of quality of fit.

These computations for the example at hand are performed below:

**Standard error of the slope $b_1$**

$$V(b_i) = \sigma^2/\Sigma(X_i - \bar{X})^2$$

$$= \sigma^2/110$$

estimate of $V(b_1) = s^2/110$

$$= 2.36/110 = 0.0215$$

estimate of standard error $(b_1) = \sqrt{\text{est. } V(b)}$

$$= 0.1465$$

Suppose $\alpha = 0.05$, so that $t\left(n-2, 1-\frac{\alpha}{2}\right)$

= t (9, 0.975) = 2.262 (from the tables of the t-distribution)

Then 95% confidence limits for $\beta_1$ are

$$b_1 \pm \frac{t(9,0.975)s}{[\Sigma(X_i - \bar{X})]^{2\frac{1}{2}}}$$

$$= 1.44 \pm (2.262)(0.1465)$$

$$= 1.44 \pm 0.3314, \text{ that is } 1.7714 \text{ and } 1.1086$$

Standard error of the Intercept $\mathbf{b_o}$

$$V(b_0) = \frac{(\Sigma x_i^2)/n}{\Sigma(X_i - \bar{X})^2}\sigma^2$$

$$= \frac{110}{11 \times 110}\sigma^2 = \sigma^2/11$$

estimate of $V(b_0) = s^2/11$

$$= \frac{2.36}{11} = 0.215$$

estimate of standard error $(b_0) = \sqrt{\text{est. } V(b_0)}$

$$= 0.4637$$

Then 95% confidence limits for $\beta_0$ are

$$b_0 \pm \frac{t(9{,}0.975)s[\Sigma x_i^2]^{\frac{1}{2}}}{[n\Sigma(x_i - \bar{x})^2]^{\frac{1}{2}}}$$

$$= 9.27 \pm (2.262)(0.4637)$$

$= 9.27 \pm 1.0489$, that is $10.3189$ and $8.2211$

**Standard error of the forecast**

Estimate of $V(\hat{Y}_k) = s^2 \left\{\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right\}$

$$= 236 \left\{\frac{1}{11} + \frac{(X_k - 0)^2}{110}\right\}$$

$$= 2.36 \left\{\frac{1}{11} + \frac{X_k^2}{110}\right\}$$

when, for instance, $x_k = \bar{X} = 0$ then

est. $V\hat{Y}_k) = 2.36 \times \frac{1}{11} = 0.214545$

$\therefore$ estimate of standard error of $\hat{\mathbf{Y}}_k = \sqrt{0.214545}$

= 0.4632

If a prediction of sales for 2009 were to be made using the regression equation

$$\hat{Y}_{\mathbf{k}} = 9.27 + 1.44X_k$$

one would obtain the value 9.27 + 1.44 (6)

= 17.91 (since the year 2009 corresponds to an X value of 6 on the transformed scale).

The 95% confidence limits for the true mean value of Y for a given $X_k$ are then given by $\hat{Y}_k \pm t(9{,}0.975)$ (est. se $Y_k$)

or $\hat{Y}_k \pm 2262 \times \left\{236\left[\frac{1}{11} + \frac{x_k^2}{110}\right]\right\}^{\frac{1}{2}}$

We shall calculate these limits for $X_k$ = 0 (year 2003) and $X_k$ = 6 (Y ear 2009)

For $X_k = 0, Y_k = 9.27$ and estimate of standard error of $\hat{Y}_k$ = 0.4632

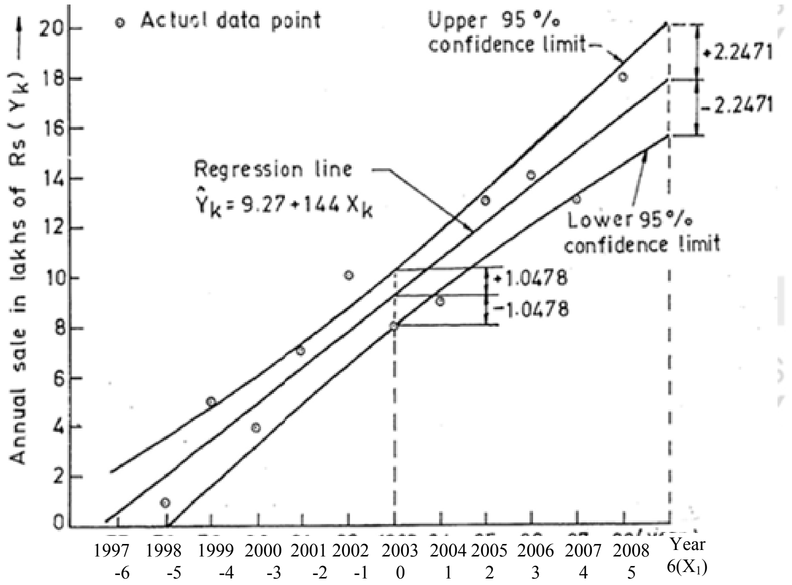
$\therefore$ 95% confidence limits are 9.27 ± (2.262 x 0.4632)

or 9.27 ± 1.0478

or 10.3178 and 8.2222

Notice that the limits, become wider as we move away from the Centre line. Figure IV illustrates the 95% confidence limits and the regression line for the example under consideration and shows how these limits change as the position of X. changes. These curves are hyperbolae. The variance and standard error of individual values may be computed by using equation (15.21), while the confidence limits for a new observation may be obtained from expression (15.22).

**Figure IV: Confidence Limits about the Regression Line**



### Activity B

For the example problem of Section 15.2 being considered above, determine the 95% and 99% confidence limits for an individual observation for a given $X_k$. Compute these limits for the year 2003 and the year 2009 (i.e. X = 0 and X = 6 respectively). How do these limits compare with those found for the mean value of Y above?

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

### F-test for Significance of Regression

From the ANOVA table constructed for the example in Section 15.2

$MS_R$= 226.95

$s^2$ = 2.36

$$F = \frac{MS_R}{s^2} = \frac{226.95}{2.36} = 9.17$$

If we look up percentage points of the F,(1,9) distribution we see that the 95% point (F1, 9, 0.95) = 5.12. Since the calculated F exceeds the critical F value in the table, that is F = 9.17 > 5.12, we reject the hypothesis $H_0$ running a risk of less than 5% of being wrong.

### Percentage Variation Explained

For the example problem $R^2 = \frac{226.95}{248.18} = 0.9145$

This indicates that the regression line explains 91.45% of the total variation about the mean.

ignou THE PEOPLE'S UNIVERSITY

## 15.5 VARIETY OF REGRESSION MODELS

The methods of regression analysis have been illustrated in this unit for the case of fitting a straight line to a giver set of data points. However the same principles are applicable to the fitting of a variety of other functions which may be relevant in certain situations highlighted below.

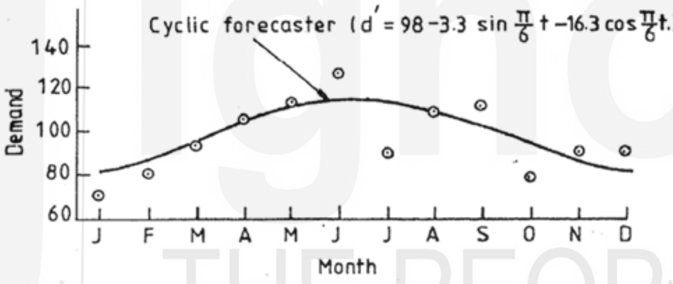
### Seasonal Model

The monthly sales for items like woollens or desert coolers is expected to be seasonal and a **sinusoidal model** would be appropriate for such a case. If $F_t$ is the forecast for period t,

$$F_1 = a + U\cos\frac{2\pi}{N}t + V\sin\frac{2\pi}{N} \ldots (15.24)$$

when a, u and v are constants, t is the time period and N is the number of time periods in a complete cycle (12 months if the cycle is 1 year). An example of such a cyclic forecaster is given in Figure V.

**Figure V: Cyclic Demand and a Cyclic Forecaster**



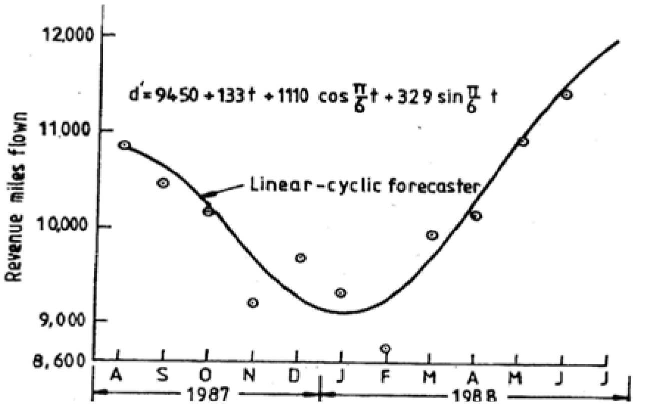
### Seasonal Models with Trend

When in addition to a cyclic component, a growth or decline over time of the demand is expected, a cyclic trend model of the following kind may be more suitable.

$$\mathrm{F_t} = a + bt + \mathrm{u}\cos\frac{2\pi}{\mathrm{N}}t + \mathrm{v}\sin\frac{2\pi}{\mathrm{N}}\mathrm{t} \ldots (15.25)$$

which is similar to equation (15.24) except for the growth term bt. Thus, there are now four parameters, a, b, u, v to be estimated. An example of such a cyclic-trend forecaster is given in Figure VI.

**Figure VI: Revenue Miles Flown and Linear-Cyclic Forecaster**



ignou
THE PEOPLE'S
UNIVERSITY

### Polynomials of Various Order

We have considered a simple model of the first order with one independent variable namely

$$Y = \beta_0 + \beta_1 X + \epsilon$$

We may have k independent variables $\mathrm{X}_1, \mathrm{X}_2 \ldots \mathrm{X}_\mathrm{k}$ and obtain a first order model with k-independent variables as

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \ldots$ (15.26)

In a forecasting context, for instance, the demand for tyres in certain month (Y) may be related to sales of petrol three months ago ($X_1$) the number of new registrations of vehicles six months ago ($X_2$) and the current months target production of vehicles ($X_3$). A second order model with one independent variable would be

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \ldots$ (15.27)

The most general type of linear model in variables $X_1, X_2, \ldots. X_k$ is of the form

$Y = \beta_0 Z_0 + \beta_1 Z_1 + \cdots \beta_p Z_p + \epsilon \ldots$ (15.28)

where

$$Z_j = Z_j(X_1, X_2 \ldots X_k)$$

can take any form: In many cases, each $Z_j$ may involve only one X variable.

### Multiplicative Models

Often by a simple transformation a non-linear model may be handled by the methods of linear regression. For instance in the multiplicative model

$$Y = \alpha X_1^b \, X_2^c \, X_3^d \qquad (15.29$$

a, b, c, d are unknown parameters and $\in$ is the multiplicative random error. Taking logarithms to the base a in equation (15.29) converts the model to the linear form

In $Y = \ln\ a + b\ln\ X_1 + C\ln\ X_2 + d\ln\ X_3 + \epsilon \ldots$ (15.30)

This model is of the form (15.28) with the parameters being In a, b, c and d and the independent variables being ln $\mathrm{X}_1$, 1n$\mathrm{X}_2$′In $\mathrm{X}_3$ While the dependent variable is 1nY.

### Linear and Non-linear Regression

We have seen above that many non-linear models can be transformed to linear models by simple transformations. It is to be noted that we are referring to linearity in the unknown parameters so that, any model which can be expressed as equation (15.28) is called linear. For such a model the parameters can be obtained by the method of least squares as the solution to a set of linear equations (known as the **normal equations**). Non-linear models which can be transformed to yield linear models are called intrinsically linear. Some models are intrinsically non-linear. Examples are:

ignou THE PEOPLE'S UNIVERSITY

$$Y = aX_1^b x_2^c X_3^d + \epsilon \qquad (15.31)$$

$$Y = \beta_0 + \beta_1^{c-\beta_2 x} + \epsilon \qquad (15.32)$$

$$Y = \beta_0 + \beta_1 X + \beta_2 (\beta_3)^x + \epsilon \qquad (15.33)$$

Some kind of interactive methods have to be employed for estimating the parameters of a non-linear system.

## 15.6 SUMMARY

In this unit fundamentals of linear regression have been highlighted. Broadly speaking, the fitting of any chosen mathematical function to given data is termed as regression analysis. The estimation of the parameters of this model is accomplished by the least squares criterion which tries to minimise the sum of squares of the errors for all the data points.

How the parameters of a fitted straight line model are estimated, has been illustrated through an example.

After the model is fitted to data the next logical question is to find out how good the quality of fit is. This question can best be answered by conducting statistical tests and determining the standard errors of estimate. This information permits us to make quantitative statements regarding confidence limits for estimates of the parameters as well as the forecast values. An overall percentage variation can also be computed and it serves to give a score to the regression. Thus it also serves to compare alternative regression models that may have been hypothesised. The various computations involved in practice have been illustrated on an example problem.

Finally, it has been emphasised that the method of least squares used in linear regression is applicable to a wide class of models. In each case the model parameters are obtained by the solution of the so called "normal equations .These are simultaneous linear equations equal in number to the number of parameters to be estimated, obtained by partially differentiating the sum of squares of errors with respect to the individual parameters.

Regression is thus a potent device for establishing relationships between variables from the given data. The discovered relationship can be used for predictive purposes. Some of the models used in forecasting of demand rely heavily on regression-analysis. One such class of models, called Time -series models is explored in Unit 20.

## 15.7 SELF-ASSESSMENT EXERCISES

1) What are the basic steps in establishing a relationship between variables from a given data?

2) What is linear regression?

In this context classify the following models as linear or non-linear.

a) $Y = a + bX + \epsilon$

ignou THE PEOPLE'S UNIVERSITY

b) $Y = a + bX + cX^2 + \epsilon$

c) $Y = a + be^{-ct} + e$

(d) $Y = a + b\cos\frac{2\pi t}{N} + c\text{Sin} + \epsilon$

3) The demand of two products A and B for twelve periods is given below:

| Period | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Demand of Produc A | 80 | 100 | 79 | 98 | 95 | 104 | 80 | 98 | 102 | 96 | 115 | 88 |
| Demand of Product B | 199 | 202 | 199 | 208 | 212 | 194 | 214 | 220 | 219 | 234 | 219 | 233 |

assuming a linear forecaster of the type $Y = \beta_0 + \beta_1 t + \in$, where Y is the demand, t the time period, $\beta_0, \beta_1$ parameters and E a random error component, establish the forecasting function for products A and B.

Obtain 95% confidence intervals for the parameters and the 95% confidence interval for the true mean value of Y at any given value of t, say $t_k$.

4) A test was run on a given process for the purpose of determining the effect of an independent variable X (such as process temperature) on a certain characteristic property of the finished product Y (such as density). Twenty observations were taken and the following results were obtained

$$\bar{X} = 5.0 \Sigma(X_i - \bar{X})^2 = 160, \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = 80$$

$$\bar{Y} = 3.0 \Sigma(Y - \bar{Y})^2 = 83.2$$

Assume a model of the type $Y = \beta_0 + \beta_1 X$ -

a) calculate the fitted regression equation

b) prepare the analysis of variable table

c) determine 95% confidence limits for the true mean value of Y when

1) X = 5.0

2) X = 9.0

5) The cost of maintenance of tractors seems to increase with the age of the tractor. The following data was collected

| **Age(yr)** | **Monthly Cost (Rs)** |
|---|---|
| 4.5 | 619 |
| 4.5 | 1049 |
| 4.5 | 1033 |
| 4.0 | 495 |
| 4.0 | 723 |

ignou THE PEOPLE'S UNIVERSITY

| | |
|---|---|
| 4.0 | 681 |
| 5.0 | 890 |
| 5.0 | 1522 |
| 5.5 | 987 |
| 5.0 | 1194 |
| 0.5 | 163 |
| 0.5 | 182 |
| 6.0 | 764 |
| 6.0 | 1373 |
| 1.0 | 978 |
| 1.0 | 466 |
| 1 0 | 549 |

Determine if a straight line relationship is sensible (use $\alpha$, the significance level = 0.10).

6) It is thought that the number of cans damaged in a box car shipment of cans is a function of the speed of the box car at impact. Thirteen box cars selected at random were used to examine whether this was true. The data collected is as follows:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Speed of ear at impact | 4 | 3 | 5 | 8 | 4 | 3 | 3 | 4 | 3 | 5 | 7 | 3 | 8 |
| No, of cans damaged | 27 | 54 | 86 | 136 | 65 | 109 | 28 | 75 | 53 | 33 | 168 | 47 | 52 |

What are your conclusions? (Use $\alpha = 0.05$)

## 15.8 KEYWORDS

**Dependent variable:** The variable of interest or focus which is influenced by one or more independent variable(s).

**Estimate:** A value obtained from data for a certain parameter of the assumed model or a forecast value obtained from the model.

**Independent variable:** A variable that can be set either to a desirable value or takes values that can be observed but not controlled. parameters of the model are estimated by minimising the sum of squares of error (discrepancy between fitted and actual value).

**Linear regression:** Fitting of any chosen mathematical model, linear in unknown parameters, to a given data.

**Model:** A general mathematical relationship relating a dependent (or response) variable Y to independent variables $X_1, X_2 \dots\dots, X_k$ by a force $Y = f(X_1, X_2 \dots X_k)$

**Non-linear regression:** Fitting-of any chosen mathematical model, non-linear in unknown parameters, to a given data.

**Parameters:** The constant terms of the chosen model that have to be estimated before the model is completely specified.

**Regression:** Relating of a dependent (or response) variable to a number independent variables, based on a given set of data.

**Response variable:** Same as a "Dependent variable".

## 15.9 FURTHER READINGS

Biegel, *J.E. Production Control -A Quantitative Approach*, Prentice Hall of India: Delhi.

Chambers, J.C., S.K. Mullick and D.D. Smith, *An Executive's Guide to Forecasting,* John Wiley: New York.

Draper, N.R. and N. Smith, *Applied Regression Analysis*, John Wiley: New York.

Firth, M., 1977. *Forecasting Methods in Business and Management*, Edward Arnold: London.

Jarrett, J., 1987. *Business Forecasting Methods,* Basil Blackwell: London.

Makridakis, S. and S.C. Wheelwright, *Interactive Forecasting*, Holden-Day: San Francisco.

Makridakis, S., S.C. Wheelwright and V.E. McGee, *Forecasting*: Methods and Applications, John Wiley: New York.

Montgomery, D.C. and L.A. Johnson, *Forecasting and Time Series Analysis*, McGiraw Hill: New York.

# UNIT 16 TIME SERIES ANALYSIS

**Objectives**

After completion of this unit, you should be able to :

- appreciate the role of time series analysis in short term forecasting
- decompose a time series into its various components
- understand auto-correlations to help identify the underlying patterns of a time series
- become aware of stochastic models developed by Box and Jenkins for time series analysis
- make forecasts from historical data using a suitable choice from available methods.

**Structure**

16.1 Introduction

16.2 Decomposition Methods

16.3 Example of Forecasting using Decomposition

16.4 Use of Auto-correlations in Identifying Time Series

16.5 An Outline of Box-Jenkins Models for Time Series

16.6 Summary

16.7 Self-assessment Exercises

16.8 Key Words

16.9 Further Readings

ignou THE PEOPLE'S UNIVERSITY

## 16.1 INTRODUCTION

Time series analysis is one of the most powerful methods in use, especially for short term forecasting purposes. From the historical data one attempts to obtain the underlying pattern so that a suitable model of the process can be developed, which is then used for purposes of forecasting or studying the internal structure of the process as a whole. We have already seen in earlier unit that a variety of methods such as subjective methods, moving averages and exponential smoothing, regression methods, causal models and time-series analysis are available for forecasting. Time series analysis looks for the dependence between values in a time series (a set of values recorded at equal time intervals) with a view to accurately identify the underlying pattern of the data.

In the case of quantitative methods of forecasting, each technique makes explicit assumptions about the underlying pattern. For instance, in using regression models we had first to make a guess on whether a linear or parabolic model should be chosen and only then could we proceed with the estimation of parameters and model-development. We could rely on mere visual inspection of the data or its graphical plot to make the best choice of the underlying model. However, such guess work, through not uncommon, is unlikely to yield very accurate or reliable results. In time series analysis, a systematic attempt is made to identify and isolate different kinds of patterns in the data. The four kinds of patterns that are most frequently encountered are horizontal, non-stationary (trend or growth), seasonal and cyclical. Generally, a random or noise component is also superimposed.

We shall first examine the method of decomposition wherein a model of the time-series in terms of these patterns can be developed. This can then be used for forecasting purposes as illustrated through an example.

A more accurate and statistically sound procedure to identify the patterns in a time-series is through the use of auto-correlations. Auto-correlation refers to the correlation between the same variable at different time lags and was discussed in Unit 18. Auto-correlations can be used to identify the patterns in a time series and suggest appropriate stochastic models for the underlying process. A brief outline of common processes and the Box-Jenkins methodology is then given.

Finally the question of the choice of a forecasting method is taken up. Characteristics of various methods are summarised along with likely situations where these may be applied. Of course, considerations of cost and accuracy desired in the forecast play a very important role in the choice.

## 16.2 DECOMPOSITION METHODS

Economic or business oriented time series are made up of four components -- trend. seasonality, cycle and randomness. Further, it is usually assumed that the relationship between these four components is multiplicative as shown in equation 16.1.

$\mathrm{X_t = T, S, C, R} \ldots$ (16.1)

where

$X_t$ is the observed value of the time series

$T_t$ denotes trend

$S_t$ denotes seasonality

$C_t$ denotes cycle

and

$R_t$ denotes randomness.

Alternatively, one could assume an additive relationship of the form

$$\mathrm{X_t = T_t + S_t + C_t + R_t}$$

But additive models are not commonly encountered in practice. We shall, therefore, be working with a model of the form (16.1) and shall systematically try to identify the individual components.

You are already familiar with the concept of moving averages, If the time series represents a seasonal pattern of L periods, then by taking a moving average of L periods, we would get the mean value for the year. Such a value will obviously be free of seasonal effects, since high months will be offset by low ones. If $M_t$ denotes the moving average of equation (16.1), it will be free of seasonality and will contain little randomness (owing to the averaging effect). Thus we can write

$M_t = T_t C_t \ldots (16.2)$

The trend and cycle components in equation (16.2) can be further decomposed by assuming some form of trend.

- One could assume different kinds of trends, such as
- linear trend, which implies a constant rate of change (Figure I)
- parabolic trend, which implies a varying rate of change (Figure II)
- exponential or logarithmic trend, which implies a constant percentage rate of change (Figure III).
- an S curve, which implies slow initial growth, with increasing rate of growth followed by a declining growth rate and eventual saturation (Figure IV).

ignou THE PEOPLE'S UNIVERSITY

**Figure I: Linear Trend**

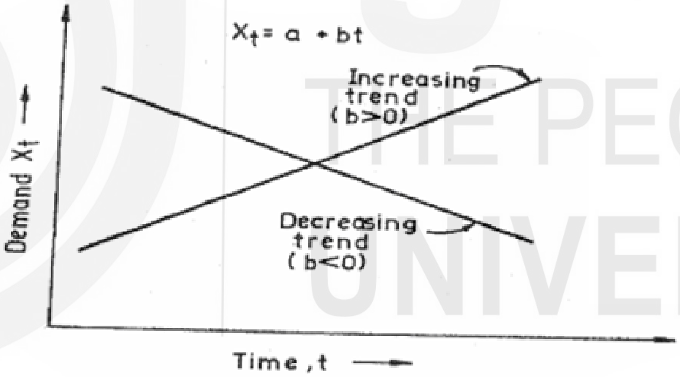


**Figure II: Parabolic Trend**

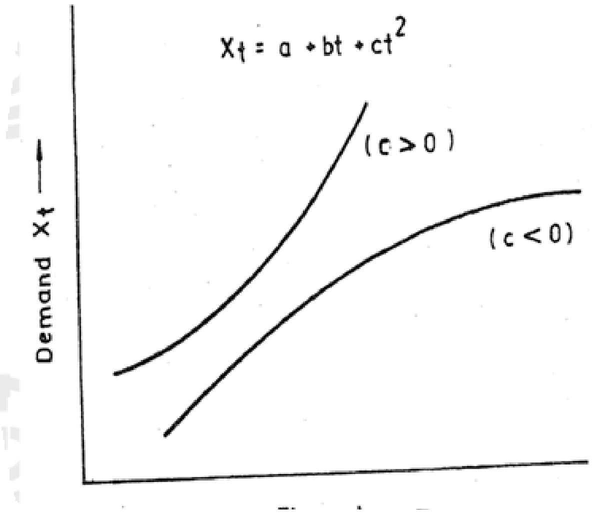


**Figure III: Exponential Trend**

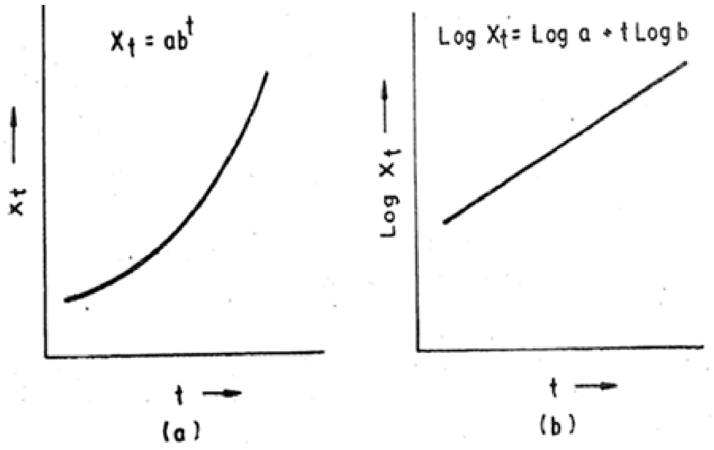
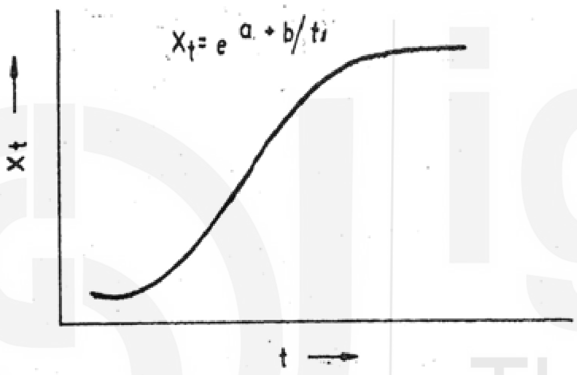


**Figure IV: A typical S Curve**



Knowing the pattern of the trend, the appropriate mathematical function could be determined from the data by using the methods of regression, as outlined in earlier unit. This would establish the values of parameters of the chosen trend model. For example, assuming a linear trend gives

$$T_t = a + bt \qquad (16.3)$$

The cycle component $C_t$ can now be isolated from the trend $T_t$ in equation (16.2) by the use of equation (16.3) as follows

$$\frac{M_t}{T_t} = \frac{M_t}{a+bt} \qquad (16.4)$$

As already indicated, if a linear trend is not adequate, one may wish to specify a non-linear one. Any pattern for the trend can be used to separate it from the cycle. In practice, however, it is often difficult to separate the two, and one may prefer to work with the trend cycle figures of equation (16.2). The isolation of the trend will add little to the overall ability to forecast. This will become clear when we take up an example problem for solution.

To isolate seasonality one could simply divide the original series (equation 16.1) by the moving average (equation 16.2) to obtain

$$\frac{X_t}{M_t} = \frac{T_t S_t C_t R_t}{T_t C_t} = S_t R_t \qquad (16.5)$$

Finally, randomness can be eliminated by averaging the different values of equation (16.5). The averaging is done on the same months or seasons of different years (for example the average of all Januaries, all Februaries,.... all Decembers). The result is a set of seasonal values free of randomness, called seasonal indices, which are widely used in practice.

In order to forecast, one must reconstruct each of the components of equation (16.1). The seasonality is known through averaging the values in equation (16.5) and the trend through (16.3). The cycle of equation (16.4) must be estimated by the user and the randomness cannot be predicted.

To illustrate the application of this procedure to actual forecasting of a time series, an example will now be considered.

## 16.3 EXAMPLE OF FORECASTING USING DECOMPOSITION

An Engineering firm producing farm equipment wants to predict future sales based on the analysis of its past sales pattern. The sales of the company for the last five years are given in Table 1.

**Table 1: Quarterly Sates of an Engineering Firm during 1983-87 (Rs. In lakhs)**

| Year | Quarters | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 1983 | 5.5 | 5.4 | 7.2 | 6.0 |
| 1984 | 4.8 | 5.6 | 6.3 | 5.6 |
| 1985 | 4.0 | 6.3 | 7.0 | 6.5 |
| 1986 | 5.2 | 6.5 | 7.5 | 7.2 |
| 1987 | 6.0 | 7.0 | 8.4 | 7.7 |

The procedure involved in the study consists of

a) deseasonalising the time series which is done by constructing a moving average M and taking the ratio $\frac{X_t}{M_t}$ which we know from equation (20.5) represents the seasonality and randomness.

b) fitting a trend line of the type T, = a + bt to the deseasonalised time series

c) identifying the cyclical variation around the trend line

d) use the above information for forecasting sales for the next year

**Deseasonalising the Time Series**

The moving averages and the ratios of the original variable to the moving average have first to the computed.

This is done in Table 2

**Table 2: Computation of moving averages $M_t$ and the ratios $X_t/M_t$**

| Year | Quarter | Actual Sales | 4 Quarter Moving Total | Centred Moving Total | Centred Moving Average ($\mathbf{M_t}$) | $\frac{\mathbf{X_1}}{\mathbf{M_1}}$ |
|---|---|---|---|---|---|---|
| 1983 | I | 5.5 | | | | |
| | II | 5.4 | | | | |
| | III | 7.2 | | 23.8 | 6.0 | 1.200 |
| | IV | 6.0 | 24.1 | 23.5 | 5.9 | 1.017 |
| 1984 | I | 4.8 | 23.4 | 23.2 | 5.8 | 0.828 |
| | II | 5.6 | 23.6 | 22.5 | 5.6 | 1.000 |
| | III | 6.3 | 22.7 | 21.9 | 5.5 | 1.145 |
| | IV | 5.6 | 22.3 | 21.9 | 5.5 | 1.018 |
| 1985 | 1 | 4.0 | 21.5 | 22.6 | 5.7 | 0.702 |
| | II | 6.3 | 22.2 | 23.4 | 5.9 | 1.068 |
| | • III | 7.0 | 22.9 | 24.4 | 6.1 | 1.148 |
| | IV | 6.5 | 23.8 | 25.1 | 6.3 | 1.032 |
| 1986 | I | 5.2 | 25.0 | 25.5 | 6.4 | 0.813 |
| | II | 6.5 | 25.2 | 26.1 | 6.5 | 1.000 |
| | III | 7.5 | 25.7 | 26.8 | 6.7 | 1.119 |
| | IV | 7.2 | 26.4 | 27.5 | 6.9 | 1.043 |
| 1987 | I | 6.0 | 27.2 | 28.2 | 7.1 | 0.845 |
| | II | 7.0 | 27.7 | 28.9 | 7.2 | 0.972 |
| | III | 8.4 | 28.6 | | | |
| | IV | 7.7 | 29.1 | | | |

ignou
THE PEOPLE'S
UNIVERSITY

It should be noticed that the 4 Quarter moving totals pertain to the middle of two successive periods. Thus the value 24.1 computed at the end of Quarter IV, 1983 refers to middle of Quarters II, III, 1983 and the next moving total of 23.4 refers to the middle of Quarters III and IV, 1983. Thus, by taking their average we obtain the centred moving total of $\frac{(24.1+23.4)}{2} = 23.75 \cong$ 23.8 to be placed for Quarter III, 1983. Similarly for the other values in case the number of periods in the moving total or average is odd, centering will not be required.

The seasonal indices for the quarterly sales data can now be computed by taking averages of the $X_t/M_t$ ratios of the respective quarters for different years as shown in Table 3.

**Table 3: Computation of Seasonal Indices**

| Year | Quarters | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 1983 | - | - | 1.200 | 1.017 |
| 1984 | 0.828 | 1.000 | 1.145 | 1.018 |
| 1985 | 0.702 | 1.068 | 1.148 | 1.032 |
| 1986 | 0.813 | 1.000 | 1.119 | 1.043 |
| 1987 | 0.845 | 0.972 | - | - |
| Mean | 0.797 | 1.010 | 1.153 | 1.028 |
| Seasonal Index | 0.799 | 1.013 | 1.156 | 1.032 |

The seasonal indices are computed from the quarter means by adjusting these values of means so that the average over the year is unity. Thus the sum of means in Table 3 is 3.988 and since there are four Quarters, each mean is adjusted by multiplying it with the constant figure of 4/3.988 to obtain the indicated seasonal indices. These seasonal indices can now be used to obtain the deseasonalised sales of the firm by dividing the actual sales by the corresponding index as shown in Table 4.

**Table 4: Deseasonalised Sales**

| Year | Quarter | Actual Sales | Seasonal index | Deseasonalised Sales |
|---|---|---|---|---|
| 1983 | I | 5.5 | 0.799 | 6.9 |
| | II | 5.4 | 1.013 | 5.3 |
| | III | 7.2 | 1.156 | 6.2 |
| | IV | 6.0 | 1.032 | 5.8 |
| 1964 | I | 4.8 | 0.799 | 6.0 |
| | II | 5.6 | 1.013 | 5.5 |
| | 111 | 6.3 | 1.156 | 5.4 |
| | IV | 5.6 | 1.032 | 5.4 |
| 1985 | 1 | 4.0 | 0.799 | 5.0 |
| | 11 | 6.3 | 1.013 | 6.2 |
| | III | 7.0 | 1.156 | 6.0 |
| | IV | 6.5 | 1.032 | 6.3 |
| 1986 | I | 5.2 | 0.799 | 6.5 |
| | II | 6.5 | 1.013 | 6.4 |
| | 111 | 7.5 | 1.156 | 6.5 |
| | IV | 7.2 | 1.032 | 7.0 |
| 1967 | 1 | 6.0 | 0.799 | 7.5 |
| | V II | 7.0 | 1.013 | 6.9 |
| | III | 8.4 | 1.156 | 7.3 |
| | IV | 7.7 | 1.032 | 7.5 |

ignou THE PEOPLE'S UNIVERSITY

### Fitting a Trend Line

The next step after deseasonalising the data is to develop the trend line. We shall here use the method of least squares that you have already studied in Unit 19 on regression. Choice of the origin in the middle of the data with a suitable scaling simplifies computations considerably. To fit a straight line of the form Y = a + bX to the deseasonalised sales, we proceed as shown in Table 5.

**Table 5: Computation of Trend**

| Year | Quarter | Deseasonalised Sales (Y) | X | $X^2$ | XY |
|---|---|---|---|---|---|
| 1983 | I | 6.9 | -19 | 361 | -131.1 |
| | II | 5.3 | -17 | 289 | -90.1 |
| | III | 6.2 | -15 | 225 | -93.0 |
| | IV | 5.8 | -13 | 169 | -75.4 |
| 1984 | I | 6.0 | -11 | 121 | -66.0 |
| | II | 5.5 | -9 | 81 | -49.5 |
| | III | 5.4 | -7 | 49 | -37.8 |
| | IV | 5.4 | -5 | 25 | -27.0 |
| 1985 | I | 5.0 | -3 | 9 | -15.0 |
| | II | 6.2 | -1 | 1 | -6.2 |
| | III | 6.0 | 1 | 1 | 6.0 |
| | IV | 6.3 | 3 | 9 | 18.9 |
| 1986 | I | 6.5 | 5 | 25 | 32.5 |
| | II | 6.4 | 7 | 49 | 44.8 |
| | III | 6.5 | 9 | 81 | 58.5 |
| | IV | 7.0 | 11 | 121 | 77.0 |
| 1987 | I | 7.5 | 13 | 169 | 97.5 |
| | II | 6.9 | 15 | 225 | 103.5 |
| | III | 7.3 | 17 | 289 | 124.1 |
| | IV | 7.5 | 19 | 361 | 142.5 |
| Total | | 125.6 | 0 | 2660 | 114.2 |

$$a = \frac{\Sigma Y}{n} = \frac{125.6}{20} = 6.3$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{114.2}{2660} = 0.04X$$

∴ the trend line is $Y = 6.3 + 0.04$X

**Identifying Cyclical Variation**

The cyclical component is identified by measuring deseasonalised variation around the trend line, as the ratio of the actual deseasonalised sales to the value predicted by the trend line. The computations are shown in Table 6.

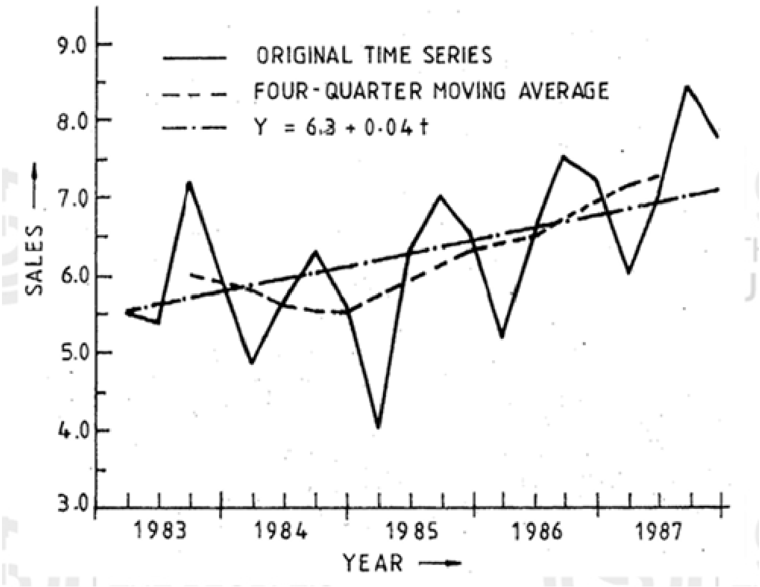
ignou
THE PEOPLE'S
UNIVERSITY

**Table 6: Computation of Cyclical Variation**

| Year | Quarter | Deseasonalised Sales (Y) | Trend a+bX | $\frac{Y}{a+bX}$ |
|---|---|---|---|---|
| 1983 | I | 6.9 | 5,54 | 1.245 |
| | II | 5.3 | 5.62 | 0.943 |
| | III | 6.2 | 5.70 | 1.088 |
| | IV | 5.8 | 5.78 | 1.003 |
| 1984 | I | 6.0 | 5.86 | 1.024 |
| | II | 5.5 | 5.94 | 0.926 |
| | III | 5.4 | 6.02 | 0.897 |
| | IV | 5.4 | 6.10 | 0.885 |
| 1985 | I | 5.0 | 6.18 | 0.809 |
| | II | 6,2 | 6.26 | 0.990 |
| | Ill | 6,0 | 6.34 | 0.946 |
| | IV | 6.3 | 6.42 | 0.981 |
| 1986 | I | 6,5 | 6.50 | 1.000 |
| | II | 6.4 | 6.58 | 0.973 |
| | III | 6,5 | 6.66 | 0.976 |
| | IV | 7.0 | 6.74 | 1.039 |
| 1987 | I | 7.5 | 6.82 | 1.110 |
| | II | 6.9 | 6.90 | 1.000 |
| | III | 7.3 | 6.98 | 1.046 |
| | IV | 7.5 | 7.06 | 1.062 |

The random or irregular variation is assumed to be relatively insignificant. We have thus described the time series in this problem using the trend, cyclical and seasonal components. Figure V represents the original time series, its four quarter moving average (containing the trend and cycle components) and the trend line.

**Figure V: Time Series with Trend and Moving Averages**



ignou THE PEOPLE'S UNIVERSITY

### Forecasting with the Decomposed Components of the Time Series

Suppose that the management of the Engineering firm is interested in estimating the sales for the second and third quarters of 1988. The estimates of the deseasohalised sales can be obtained by using the trend line

Y = 6.3 + 0.04(23)

= 7.22 (2nd Quarter 1988)

and Y = 6.3 + 0.04 (25)

= 7.30 (3rd Quarter 1988)

These estimates will now have to be seasonalised for the second and third quarters respectively. This can be done as follows :

For 1988 2nd quarter

seasonalised sales estimate = 7.22 x 1.013 = 7.31

For 1988 3rd quarter

seasonalised sales estimate = 7.30 x 1.56

= 8.44

Thus, on the basis of the above analysis, the sales estimates of the Engineering firm for the second and third quarters of 1988 are Rs. 7.31 lakh and Rs. 8.44 lakh respectively.

These estimates have been obtained by taking the trend and seasonal variations into account. Cyclical and irregular components have not been taken into account. The procedure for cyclical variations only helps to study past behaviour and does not help in predicting the future behaviour.

Moreover, random or irregular variations are difficult to quantify.

## 16.4 USE OF AUTO-CORRELATIONS IN IDENTIFYING TIME SERIES

While studying correlation in Unit 14, auto-correlation was defined as the correlation of a variable with itself, but with a time lag. The study of auto-correlation provides very valuable clues to the underlying pattern of a time series. It can also be used to estimate the length of the season for seasonality. (Recall that in the example problem considered in the previous. section, we assumed that a complete season consisted of four quarters.)

When the underlying time series represents completely random data, then the graph of auto-correlations for various time lags stays close to zero with values fluctuating both on the +ve and -ve side but staying within the control limits. This in fact represents a very convenient method of identifying randomness in the data.

If the auto-correlations drop slowly to zero, and more than two or three differ significantly from zero, it indicates the presence of a trend in the data. This trend can be-removed by differentiating (that is taking differences between consecutive values and constructing a new series).

A seasonal pattern in the data would result in the auto-correlations oscillating around zero with some values differing significantly from zero. The length of seasonality can be determined either from the number of periods it takes for the auto-correlations to make a complete cycle or by the time lag giving the largest auto correlation.

For any given data, the plot of auto-correlation for various time lags is diagnosed to identify which of the above basic patterns (or a combination of these patterns) it follows. This is broadly how auto-correlations are used to identify the structure of the underlying model to be chosen. The underlying mathematics and computational burden tend to be heavy and involved. Computer routines for carrying out computations are available. The interested reader may refer to Makridakis and Wheelwright for further details.

## 16.5 AN OUTLINE OF BOX-JENKINS MODELS FOR TIME SERIES

Box and Jenkins (1976) have proposed a sophisticated methodology for stochastic model building and forecasting using time series. The purpose of this section is merely to acquaint you with some of the terms, models and methodology developed by Box and Jenkins.

A time series may be classified as stationary (in equilibrium about a constant mean value) or non-stationary (when the process has no natural or stable mean). In stochastic model building the non-stationary processes often converted to a stationary one by differencing. The two major classes of models used popularly in time series analysis are Auto-regressive and Moving Average models.

**Auto-regressive Models**

In such models, the current value of the process is expressed as a finite, linear aggregate of previous values of the process and a random shock or error $a_t$. Let us denote the value of a process at equally spaced times $t, t-1, t-2$ ... by $Z_t, Z_{t-1}, Z_{t-2}$ ...... also let $Z_t$, $Z_{t-1}$ $Z_{t-2}$ ... be the deviations from the process mean, m. That is $\bar{Z}_t = Z_t - m$. Then

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t\cdot p} + a_t \ \ldots \ (16.6)$$

is called an auto-regressive (AR) process of order p. The reason for this name is that equation (16.6) represents a regression of the variable $Z_t$ on successive values of itself. The model contains p + 2 unknown parameters m, $\phi_1, \phi_2, \ldots\ldots \phi_p, \sigma_a^2$ which in practice have to be estimated from the data.

The additional parameter $\sigma^2$ is the variance of the random error component.

**Moving Average models**

Another kind of model of great importance is the moving average model where $Z_t$ is made linearly dependent on a finite number q of previous a's (error terms)

Thus

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_q a_{t-q} \ \ldots \qquad (16.7)$$

ignou
THE PEOPLE'S
UNIVERSITY

is called a moving average (MA) process of order q. The name "moving average" is somewhat misleading because the weights $1, -\theta_1, -\theta_2, \ldots, -\theta_q$ which multiply the a's, need not total unity nor need they be positive. However, this nomenclature is in common use and therefore we employ it. The model (16.7) contains q + 2 unknown parameters m, $\theta_1, \theta_2, \ldots . \theta_q, \sigma_a^2$ which in practice have to be estimated from the data.

**Mixed Auto-regressive-moving average models :**

It is sometimes advantageous to include both auto-regressive and moving average terms in the model. This leads to the mixed auto-regressive-moving average (ARMA) model.

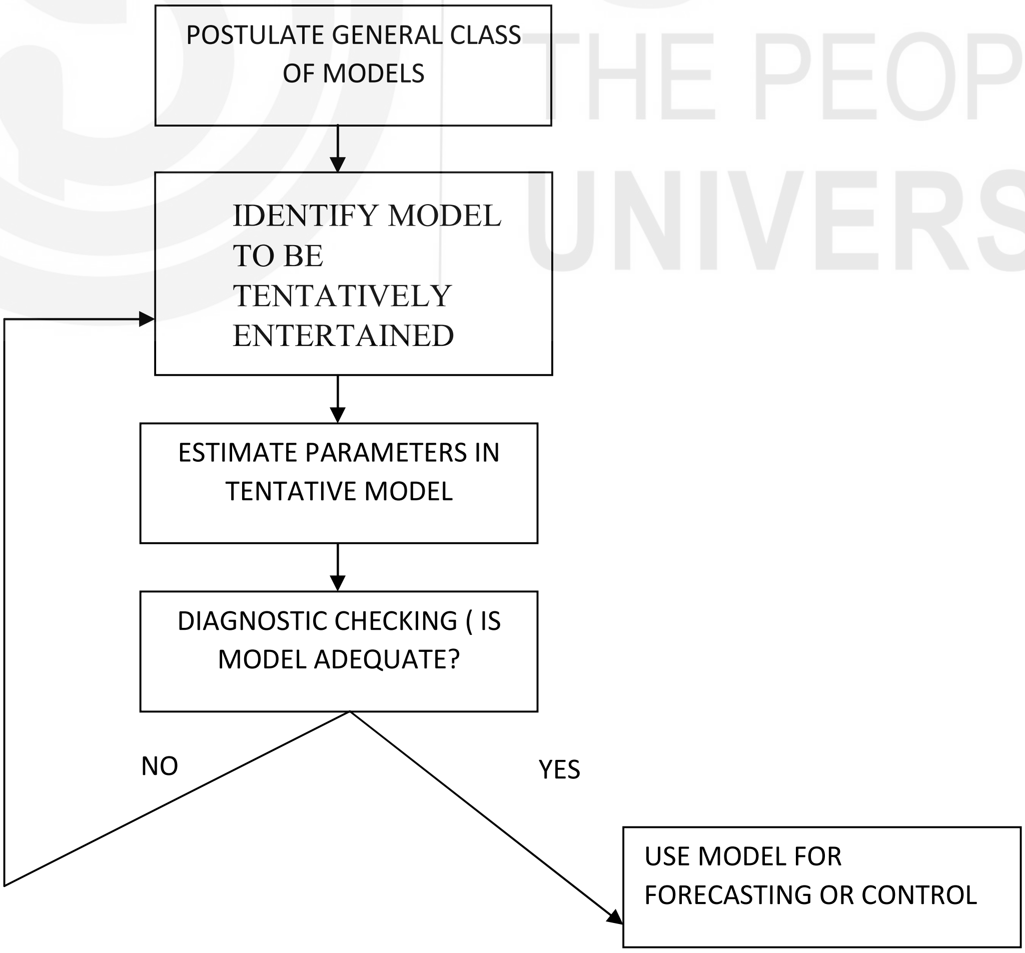
$$Z_1 = \phi_1 Z_{t-1} + \cdots \phi_p Z_{t-p} + a_t - \theta_i a_{t-1} - \cdots - \theta_q a_{l-q} \ldots (16.8)$$

In using such models in practice p and q are not greater than 2.

For non-stationary processes the most general model used is an auto-regressive integrated moving average (ARIMA) process of order (p, d, q) where d represents the degree of differencing to achieve stationarity in the process.

The main contribution of Box and Jenkins is the development of procedures for identifying the ARMA model that best fits a set of data and for testing the adequacy of that model. The various stages identified by Box and Jenkins in their interactive approach to model building are shown in Figure VI. For details on how such models are developed refer to Box and Jenkins.

**Figure VI: The Box-Jenkins Methodology**


POSTULATE GENERAL CLASS OF MODELS
IDENTIFY MODEL TO BE TENTATIVELY ENTERTAINED
ESTIMATE PARAMETERS IN TENTATIVE MODEL
DIAGNOSTIC CHECKING ( IS MODEL ADEQUATE?
NO
YES
USE MODEL FOR FORECASTING OR CONTROL

ignou
THE PEOPLE'S
UNIVERSITY

## 16.6 SUMMARY

Some procedures for time series analysis have been described in this unit with a view to making more accurate and reliable forecasts of the future. Quite often the question that puzzles a person is how to select an appropriate forecasting method. Many times the problem context or time horizon involved would decide the method or limit the choice of methods. For instance, in new areas of technology forecasting where historical information is scanty, one would resort, to some subjective method like opinion poll or a DELPHI study. In situations where one is trying to control or manipulate a factor a causal model might be appropriate in identifying the key variables and their effect on the dependent variable.

In this particular unit, however, time series models or those models where historical data on demand or the variable of interest is available are discussed. Thus we are dealing with projecting into the future from the past. Such models are short term forecasting models.

The decomposition method has been discussed. Here the time series is broken up into seasonal, trend, cycle and random components from the given data and reconstructed for forecasting purposes. A detailed example to illustrate the procedure is also given.

Finally the framework of stochastic models used by Box and Jenkins for time series analysis has been outlined. The AR, MA, ARMA and ARIMA processes in Box- Jenkins models are briefly described so that the interested reader can pursue a detailed study on his own.

ignou THE PEOPLE'S UNIVERSITY

## 16.7 SELF-ASSESSMENT EXERCISES

1) What do you understand by time series analysis? How would you go about conducting such an analysis for forecasting the sales of a product in your firm?
2) Compare time series analysis with other methods of forecasting, briefly summarising the strengths and weaknesses of various methods.
3) What would be the considerations in the choice of a forecasting method?
4) Find the 4-quarter moving average of the following time series representing the quarterly production of coffee in an Indian State.

| | Production (in Tonnes) | | | |
|---|---|---|---|---|
| **Year** | **Quarter I** | **Quarter II** | **Quarter III** | **Quarter IV** |
| 1983 | 5 | 1 | 10 | 17 |
| 1984 | 7 | 1 | 10 | 16 |
| 1985 | 9 | 3 | 8 | 18 |
| 1986 | 5. | 2 | 15 | 19 |
| 1987 | 8 | 4 | 14 | 21 |

5) Given below is the data of production of a certain company in lakhs of units

| Year | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 |
|---|---|---|---|---|---|---|---|
| Production | 15 | 14 | 18 | 20 | 17 | 24 | 27 |

a) Compute the linear trend by the method of least squares.

b) Compute the trend values of each of the years.

6) Given the following data on factory production of a certain brand of motor vehicles, determine the seasonal indices by the ratio to moving average method for August and September, 1985.

| | Production (in thousand units) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Jan. | Feb. | Mar. | Apr. | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
| 1985 | 7.92 | 7.81 | 7.91 | 7.03 | 7.25 | 7.17 | 5.01 | 3.90 | 4.64 | 7.03 | 6.88 | 6.14 |
| 1986 | 4.86 | 4.48 | 5.26 | 5.48 | 6.42 | 6.82 | 4.98 | 2.45 | 4.51 | 6.38 | 6.38 | 7.59 |

7) A survey of used car sales in a city for the 10-year period 1976-85 has been made. A linear trend was fitted to the sales for month for each year and the equation was found to be

$Y = 400 + 18\ t$

where t = 0 on January 1, 1981 and t is measured in $\frac{1}{2}$ year (6 monthly) units

a) use this trend to predict sales for June, 1990

b) If the actual sales in June. 1987 are 600 and the relative seasonal index June sales is 1.20, what would be the relative cyclical, irregular index June, 1987?

8) The monthly sales for the last one year of a product in thousands of units are given below :

| **Month** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sales | 0.5 | 1.5 | 2.2 | 3.0 | 3.2 | 3.5 | 3.5 | 3.5 | 3.8 | 4.0 | 4.7 | 5.5 |

Compute the auto-correlation coefficients up to lag 4. What conclusion can be derived from these values regarding the presence of a trend in the data?

## 16.8 KEY WORDS

**Auto-correiation :** Similar to correlation in that it Describes the association between values of the same variable but at different time periods. Auto-corre'aHo" coefficients provide important information about the underlying patterns in the data.

**Auto-regressive/Moving Average (ARMA) Models :** Auto-regressive(AR) models assume that future values are linear combinations of past values. Moving Average (MA) models, on the other hand, assume that future values are linear combinations of past errors. .A combination of the two is called an "Auto-regressive/Moving Average (ARMA) model".

**Decomposition :** Identifying the trend, seasonality, cycle and randomness in a time series.

**Forecasting :** Predicting the future values of a variable based on historical values of the same or other variable(s). If the forecast is based simply on past values of the variable itself, it is called time series forecasting, otherwise it is a causal type forecasting.

**Seasonal Index :** A number with a base of 1.00 that indicates the seasonality for a given period in relation to other periods.

**Time Series Model :** A model that predicts the future by expressing it as a function of the past.

**Trend :** A growth or decline in the mean value of a variable over the relevant time span.

## 16.9 FURTHER READINGS

Box, G.E.P. and G.M. *Jenkinsx. Time Series Analysis, Forecasting and Control*, Holden-Day: San Francisco.

Chambers, J.C., S.K. Mullick and D.D. Smith. *An Executive's Guide to Forecasting*, John Wiley: New York.

Makridakis, S. and S. Wheelwright. *interactive Forecasting: Univariate and Multivariate Methods,* Holden-Day: San Francisco.

Makridakis, S. and S. Wheelwright. *Forecasting: Methods and Applications*, John Wiley, New York.

Montgomery, D.C. and L.A. Johnson. *Forecasting and Time Series Analysis,* McGraw Hill: New York.

Nelson, C.R., *Applied Time Series Analysis for Managerial Forecasting,* Holden-Day: